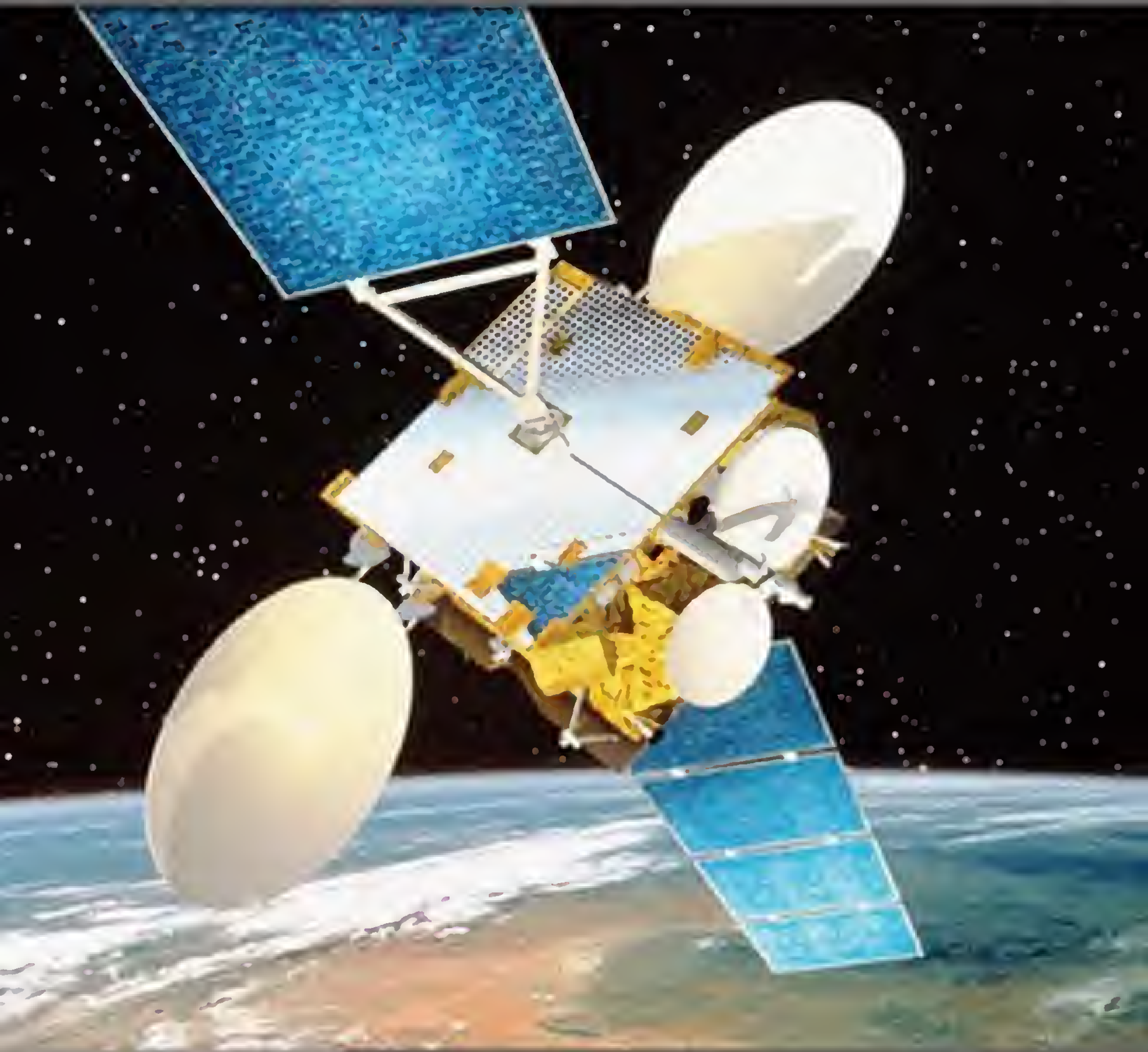
SATELLITE COMMUNICATIONS

Fourth Edition



DENNIS RODDY

- Mobile services
- Digital television services
- Satellite-based Internet and ATM services

Satellite Communications

Dennis Roddy

Fourth Edition

McGraw-Hill
New York Chicago San Francisco Lisbon London Madrid Mexico City Milan New Delhi San Juan Seoul Singapore Sydney Toronto

The McGraw-Hill Companies

CIP Data is on file with the Library of Congress

1 2 3 4 5 6 7 8 9 0 DOC/DOC 0 1 2 1 0 9 8 7 6

ISBN 0-07-146298-8

The sponsoring editor for this book was Stephen S. Chapman and the production supervisor was Richard C. Ruzycka. It was set in Century Schoolbook by International Typesetting and Composition. The art director for the cover was Anthony Landi.

Printed and bound by RR Donnelley.

The first edition of this book was published by Prentice-Hall Inc., copyright © 1989.



This book is printed on recycled, acid-free paper containing a minimum of 50% recycled, de-inked fiber.

Contents

Preface

As with previous editions, the fourth edition provides broad coverage of satellite communications, while maintaining sufficient depth to lay the foundations for more advanced studies. Mathematics is used as a descriptive tool and to obtain numerical results, but lengthy mathematical derivations are avoided. In setting up numerical problems and examples the author has made use of Mathcad™, a computer program, but the worked examples in the text are presented in normal algebraic notation, so that other programs, including programmable calculators can be used.

The main changes compared to the previous edition are as follows. In Chap. 1 the sections on INTELSAT and polar orbiting satellites, including environmental and search and rescue, have been updated. Sun-synchronous orbits have been treated in more detail in Chap. 2. A new section on planar antennas and arrays, including reflectarrays, and array switching has been added in Chap. 6. Chapter 8 includes additional details on C-band reception of television signals. In Chap. 11, a more detailed description is given of linear block codes, and new sections on the Shannon capacity and *turbo and low density parity check* (LDPC) codes have been included. Chapter 12 has a new section on *intersatellite links* (ISLs), including optical links. Chapter 15 has been extensively rewritten to include more basic details on networks and *asynchronous transfer mode* (ATM) operation. Chapter 16 covers *high definition television* (HDTV) in more detail, and the Iridium mobile satellite system, which is now under new ownership, is described in Chap. 17.

In this age of heightened security concerns, it has proved difficult to get detailed technical information on satellite systems and equipment. Special thanks are, therefore, due to the following people and organizations that provided copies of technical papers, diagrams, and figures for the topics listed:

Planar antennas and arrays, reflectarrays, and array switching: Jacquelyn Adams, Battelle/GLITeC; Dr. Luigi Boccia, Universita della

Calabria; Thomas J. Braviak, Director, Marketing Administration, Aeroflex/KDI-Integrated Products; Dr. Michael Parnes, Ascor, Saint-Petersburg, Russia; Dharmesh Patel, Radar Division Naval Research Laboratory; Professor David Pozar, Electrical and Computer Engineering, University of Massachusetts at Amherst; Dr. Bob Romanofsky, Antenna, Microwave, and Optical Systems Branch, NASA Glenn Research Center; Dr. Peter Schrock, Webmaster/Publications Representative, JPL.

ATM: William D. Ivancic, Glenn Research Center, and Lewis Research Center, NASA; Dr.-Ing. Petia Todorova, Fraunhofer Institut FOKUS, Berlin.

Turbo and LDPC codes: Dr. Alister G. Burr, Professor of Communications, Dept. of Electronics, University of York; Tony Summers, Senior Applications Engineer, Comtech AHA.

HDTV: Cathy Firebrace, Information Officer, the IEE, London U.K.

INTELSAT: Travis S. Taylor, Corporate Communications, Intelsat, Washington, DC.

Iridium system: Liz DeCastro, Corporate Communications Director, Iridium Satellite.

Cospas-Sarsat: Cheryl Bertoia, Principal Operations Officer, Deputy Head, Cospas-Sarsat Secretariat, London, U.K., and Hannah Bermudez also of the Cospas-Sarsat Secretariat.

And from the third edition, thanks to Dr. Henry Driver of Computer Sciences Corporation for details relating to the calculation of geodetic position in Chap. 2. Thanks also to the users and reviewers who made suggestions for corrections, additions, and improvements. Errors should not occur, but they do, and the author would be grateful if these are drawn to his attention. He can be reached at droddy@tbayel.com.

The editorial team at McGraw-Hill has contributed a great deal in getting the fourth edition into print: thanks are due to Steve Chapman, the sponsoring editor; Diana Mattingly, editorial assistant; and Gita Raman, project manager, for their help, and their gentle but persistent reminders to keep the book on schedule.

Dennis Roddy
Thunder Bay, Ontario

Chapter

1

Overview of Satellite Systems

1.1 Introduction

The use of satellites in communications systems is very much a fact of everyday life, as is evidenced by the many homes equipped with antennas, or "dishes," used for reception of satellite television. What may not be so well known is that satellites form an essential part of telecommunications systems worldwide, carrying large amounts of data and telephone traffic in addition to television signals.

Satellites offer a number of features not readily available with other means of communications. Because very large areas of the earth are visible from a satellite, the satellite can form the star point of a communications net, simultaneously linking many users who may be widely separated geographically. The same feature enables satellites to provide communications links to remote communities in sparsely populated areas that are difficult to access by other means. Of course, satellite signals ignore political boundaries as well as geographic ones, which may or may not be a desirable feature.

To give some idea of cost, the construction and launch cost of the Canadian Anik-E1 satellite (in 1994 Canadian dollars) was $281.2 million, and that of the Anik-E2, $290.5 million. The combined launch insurance for both satellites was $95.5 million. A feature of any satellite system is that the cost is *distance insensitive,* meaning that it costs about the same to provide a satellite communications link over a short distance as it does over a large distance. Thus a satellite communications system is economical only where the system is in continuous use and the costs can be reasonably spread over a large number of users.

Satellites are also used for remote sensing, examples being the detection of water pollution and the monitoring and reporting of

weather conditions. Some of these remote sensing satellites also form a vital link in search and rescue operations for downed aircraft and the like.

A good overview of the role of satellites is given by Pritchard (1984) and Brown (1981). To provide a general overview of satellite systems here, three different types of applications are briefly described in this chapter: (1) the largest international system, Intelsat, (2) the domestic satellite system in the United States, Domsat, and (3) U.S. *National Oceanographic and Atmospheric Administration* (NOAA) series of polar orbiting satellites used for environmental monitoring and search and rescue.

1.2 Frequency Allocations for Satellite Services

Allocating frequencies to satellite services is a complicated process which requires international coordination and planning. This is carried out under the auspices of the *International Telecommunication Union* (ITU).

To facilitate frequency planning, the world is divided into three regions:

Region 1: Europe, Africa, what was formerly the Soviet Union, and Mongolia

Region 2: North and South America and Greenland

Region 3: Asia (excluding region 1 areas), Australia, and the southwest Pacific

Within these regions, frequency bands are allocated to various satellite services, although a given service may be allocated different frequency bands in different regions. Some of the services provided by satellites are:

Fixed satellite service (FSS)
Broadcasting satellite service (BSS)
Mobile satellite services
Navigational satellite services
Meteorological satellite services

There are many subdivisions within these broad classifications; for example, the FSS provides links for existing telephone networks as well as for transmitting television signals to cable companies for distribution over cable systems. Broadcasting satellite services are intended mainly for direct broadcast to the home, sometimes referred

to as *direct broadcast satellite* (DBS) service [in Europe it may be known as *direct-to-home* (DTH) service]. Mobile satellite services would include land mobile, maritime mobile, and aeronautical mobile. Navigational satellite services include *global positioning systems* (GPS), and satellites intended for the meteorological services often provide a search and rescue service.

Table 1.1 lists the frequency band designations in common use for satellite services. The Ku band signifies the band under the K band, and the Ka band is the band above the K band. The Ku band is the one used at present for DBS, and it is also used for certain FSS. The C band is used for FSS, and no DBS is allowed in this band. The very high frequency (VHF) band is used for certain mobile and navigational services and for data transfer from weather satellites. The L band is used for mobile satellite services and navigation systems. For the FSS in the C band, the most widely used subrange is approximately 4 to 6 GHz. The higher frequency is nearly always used for the uplink to the satellite, for reasons that will be explained later, and common practice is to denote the C band by 6/4 GHz, giving the uplink frequency first. For the direct broadcast service in the Ku band, the most widely used range is approximately 12 to 14 GHz, which is denoted by 14/12 GHz. Although frequency assignments are made much more precisely, and they may lie somewhat outside the values quoted here (an example of assigned frequencies in the Ku band is 14,030 and 11,730 MHz), the approximate values stated are quite satisfactory for use in calculations involving frequency, as will be shown later in the text.

Care must be exercised when using published references to frequency bands, because the designations have been developed somewhat differently for radar and communications applications; in addition, not all countries use the same designations.

TABLE 1.1 Frequency Band Designations

Frequency range, (GHz)	Band designation
0.1–0.3	VHF
0.3–1.0	UHF
1.0–2.0	L
2.0–4.0	S
4.0–8.0	C
8.0–12.0	X
12.0–18.0	Ku
18.0–27.0	K
27.0–40.0	Ka
40.0–75	V
75–110	W
110–300	mm
300–3000	μm

TABLE 1.2 ITU Frequency Band Designations

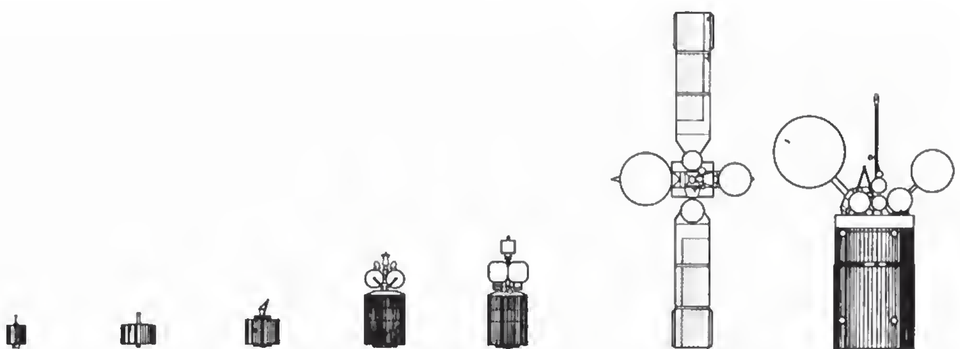
Band number	Symbols	Frequency range (lower limit exclusive, upper limit inclusive)	Corresponding metric subdivision	Metric abbreviations for the bands
4	VLF	3–30 kHz	Myriametric waves	B.Mam
5	LF	30–300 kHz	Kilometric waves	B.km
6	MF	300–3000 kHz	Hectometric waves	B.hm
7	HF	3–30 MHz	Decametric waves	B.dam
8	VHF	30–300 MHz	Metric waves	B.m
9	UHF	300–3000 MHz	Decimetric waves	B.dm
10	SHF	3–30 GHz	Centimetric waves	B.cm
11	EHF	30–300 GHz	Millimetric waves	B.mm
12		300–3000 GHz	Decimillimetric waves	

SOURCE: ITU Geneva.

The official ITU frequency band designations are shown in Table 1.2 for completeness. However, in this text the designations given in Table 1.1 will be used, along with 6/4 GHz for the C band and 14/12 GHz for the Ku band.

1.3 INTELSAT

INTELSAT stands for *International Telecommunications Satellite.* The organization was created in 1964 and currently has over 140 member countries and more than 40 investing entities (see http://www.intelsat.com/ for more details). In July 2001 INTELSAT became a private company and in May 2002 the company began providing end-to-end solutions through a network of teleports, leased fiber, and *points of presence* (PoPs) around the globe. Starting with the Early Bird satellite in 1965, a succession of satellites has been launched at intervals of a few years. Figure 1.1 illustrates the evolution of some of the INTELSAT satellites. As the figure shows, the capacity, in terms of number of voice channels, increased dramatically with each succeeding launch, as well as the design lifetime. These satellites are in *geostationary orbit,* meaning that they appear to be stationary in relation to the earth. The geostationary orbit is the topic of Chap. 3. At this point it may be noted that geostationary satellites orbit in the earth's equatorial plane and their position is specified by their longitude. For international traffic, INTELSAT covers three main regions—the *Atlantic Ocean Region* (AOR), the *Indian Ocean Region* (IOR), and the *Pacific Ocean Region* (POR) and what is termed *Intelsat America's Region.* For the ocean regions the satellites are positioned in geostationary orbit above the particular ocean, where they provide a transoceanic telecommunications route. For example, INTELSAT satellite 905 is positioned at 335.5° east longitude. The footprints for the C-band antennas are shown in Fig. 1.2*a*, and for the Ku-band spot beam antennas in Figs. 1.2*b* and *c*.



Designation: Intelsat	I	II	III	IV	IV A	V	V A/V B	VI
Year of first launch	1965	1966	1968	1971	1975	1980	1984/85	1986/87
Prime contractor	Hughes	Hughes	TRW	Hughes	Hughes	Ford Aerospace	Ford Aerospace	Hughes
Width (m)	0.7	1.4	1.4	2.4	2.4	2.0	2.0	3.6
Height (m)	0.8	0.7	1.0	5.3	6.8	6.4	6.4	6.4
Launch vehicles		Thor Delta		Atlas-Centaur		Atlas-Centaur end Ariane	Atlas-Centaur and Ariane	STS end Ariane
Spacecraft mass in transfer orbit (kg)	68	182	293	1385	1489	1946	2140	12,100/3720
Communications payload mass (kg)	13	36	56	185	190	235	280	800
End-of-life (EOL) power of equinox (W)	40	75	134	480	800	1270	1270	2200
Design lifetime (years)	1.5	3	5	7	7	7	7	10
Capacity (number of voice channels)	480	480	2400	8000	12,000	25,000	30,000	80,000
Bandwidth (MHz)	50	130	300	500	800	2137	2480	3520

Figure 1.1 Evolution of INTELSAT satellites. (*From Colino 1985; courtesy of ITU Telecommunications Journal.*)





Figure 1.2 INTELSAT satellite 905 is positioned at 335.5° E longitude. (*a*) The footprints for the C-band antennas; (*b*) the Ku-band spot 1 beam antennas; and (*c*) the Ku-band spot 2 beam antennas.



Figure 1.2 (*Continued*).

The INTELSAT VII-VII/A series was launched over a period from October 1993 to June 1996. The construction is similar to that for the V and VA/VB series, shown in Fig. 1.1, in that the VII series has solar sails rather than a cylindrical body. This type of construction is described in more detail in Chap. 7. The VII series was planned for service in the POR and also for some of the less demanding services in the AOR. The antenna beam coverage is appropriate for that of the POR. Figure 1.3 shows the antenna beam footprints for the C-band hemispheric coverage and zone coverage, as well as the spot beam coverage possible with the Ku-band antennas (Lilly, 1990; Sachdev et al., 1990). When used in the AOR, the VII series satellite is inverted north for south (Lilly, 1990), minor adjustments then being needed only to optimize the antenna patterns for this region. The lifetime of these satellites ranges from 10 to 15 years depending on the launch vehicle. Recent figures from the INTELSAT Web site give the capacity for the INTELSAT VII as 18,000 two-way telephone circuits and three TV channels; up to 90,000 two-way telephone circuits can be achieved with the use of "digital circuit multiplication." The INTELSAT VII/A has a capacity of 22,500 two-way telephone circuits and three TV channels; up to 112,500 two-way telephone circuits can be achieved with the use of digital circuit multiplication. As of May 1999, four satellites were in service over the AOR, one in the IOR, and two in the POR.

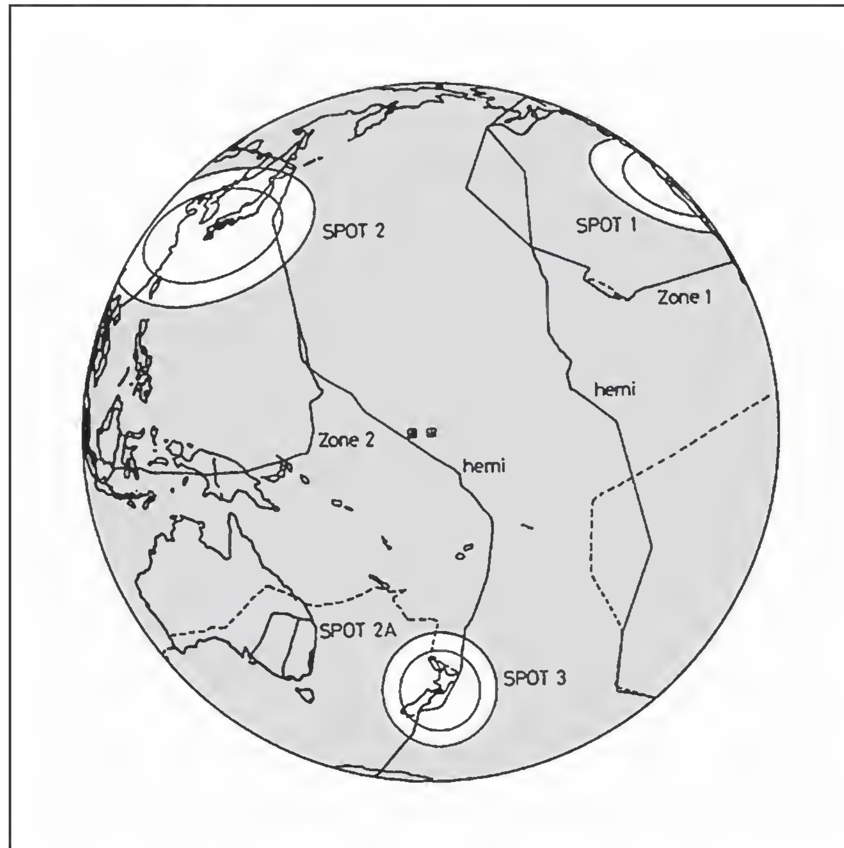


Figure 1.3 INTELSAT VII coverage (Pacific Ocean Region; global, hemispheric, and spot beams). (*From Lilly, 1990, with permission.*)

The INTELSAT VIII-VII/A series of satellites was launched over the period February 1997 to June 1998. Satellites in this series have similar capacity as the VII/A series, and the lifetime is 14 to 17 years.

It is standard practice to have a spare satellite in orbit on high-reliability routes (which can carry preemptible traffic) and to have a ground spare in case of launch failure. Thus the cost for large international schemes can be high; for example, series IX, described later, represents a total investment of approximately $1 billion.

Table 1.3 summarizes the details of some of the more recent of the INTELSAT satellites. These satellites provide a much wider range of services than those available previously, including such services as Internet, DTH TV, tele-medicine, tele-education, and interactive video and multimedia. Transponders and the types of signals they carry are

TABLE 1.3 INTELSAT Geostationary Satellites

Satellite	Location	Number of transponders	Launch date
901	342°E	Up to 72 @ 36 MHz in C-Band Up to 27 @ 36 MHz in Ku Band	June 2001
902	62°E	Up to 72 @ 36 MHz in C-Band Up to 23 @ 36 MHz in Ku Band	August 2001
903	325.5°E	Up to 72 @ 36 MHz in C-Band Up to 22 @ 36 MHz in Ku Band	March 2002
904	60°E	Up to 72 @ 36 MHz in C-Band Up to 22 @ 36 MHz in Ku Band	February 2002
905	335.5°E	Up to 72 @ 36 MHz in C-Band Up to 22 @ 36 MHz in Ku Band	June 2002
906	64°E	Up to 72 @ 36 MHz in C-Band Up to 22 @ 36 MHz in Ku Band	September 2002
907	332.5°E	Up to 72 @ 36 MHz in C-Band Up to 23 @ 36 MHz in Ku Band	February 2003
10-02	359°E	Up to 70 @ 36 MHz in C-Band Up to 36 @ 36 MHz in Ku Band	June 2004

described in detail in later chapters, but for comparison purposes it may be noted that one 36 MHz transponder is capable of carrying about 9000 voice channels, or two analog TV channels, or about eight digital TV channels.

In addition to providing transoceanic routes, the INTELSAT satellites are also used for domestic services within any given country and regional services between countries. Two such services are Vista for telephone and Intelnet for data exchange. Figure 1.4 shows typical Vista applications.

1.4 U.S. Domsats

Domsat is an abbreviation for *domestic satellite*. Domestic satellites are used to provide various telecommunications services, such as voice, data, and video transmissions, within a country. In the United States, all domsats are situated in geostationary orbit. As is well known, they make available a wide selection of TV channels for the home entertainment market, in addition to carrying a large amount of commercial telecommunications traffic.

U.S. Domsats, which provide a DTH television service, can be classified broadly as high power, medium power, and low power (Reinhart, 1990). The defining characteristics of these categories are shown in Table 1.4.

The main distinguishing feature of these categories is the *equivalent isotropic radiated power* (EIRP). This is explained in more detail in Chap. 12, but for present purposes it should be noted that the upper limit

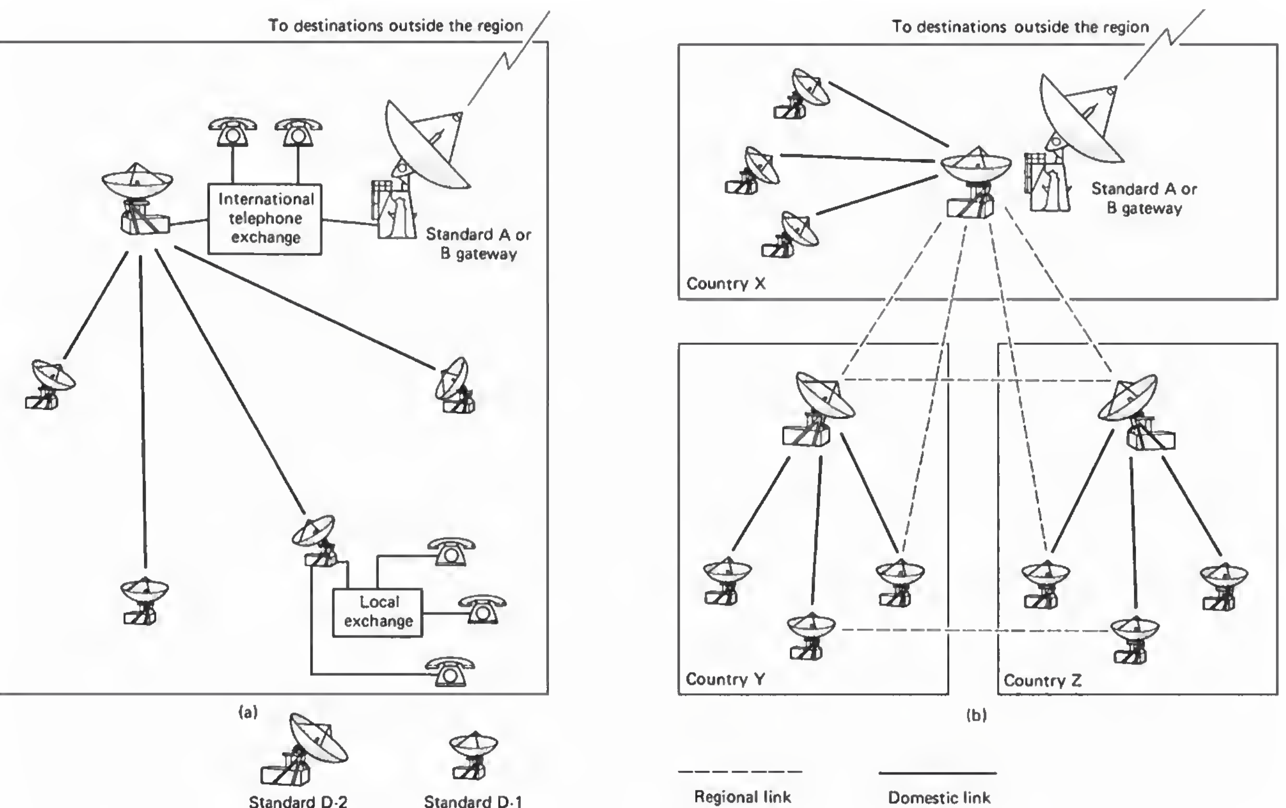


Figure 1.4 (*a*) Typical Vista application; (*b*) domestic/regional Vista network with standard A or B gateway. (*From Colino, 1985; courtesy of ITU Telecommunication Journal.*)

TABLE 1.4 Defining Characteristics of Three Categories of United States DBS Systems

	High power	Medium power	Low power
Band	Ku	Ku	C
Downlink frequency allocation GHz	12.2–12.7	11.7–12.2	3.7–4.2
Uplink frequency allocation GHz	17.3–17.8	14–14.5	5.925–6.425
Space service	BSS	FSS	FSS
Primary intended use	DBS	Point-to-point	Point-to-point
Allowed additional use	Point-to-point	DBS	DBS
Terrestrial interference possible	No	No	Yes
Satellite spacing degrees	9	2	2–3
Satellite spacing determined by	ITU	FCC	FCC
Adjacent satellite interference possible?	No	Yes	Yes
Satellite EIRP range (dBW)	51–60	40–48	33–37

NOTES: ITU—International Telecommunication Union; FCC—Federal Communications Commission.
SOURCE: Reinhart, 1990.

of EIRP is 60 dBW for the high-power category and 37 dBW for the low-power category, a difference of 23 dB. This represents an increase in received power of $10^{2.3}$ or about 200:1 in the high-power category, which allows much smaller antennas to be used with the receiver. As noted in the table, the primary purpose of satellites in the high-power category is to provide a DBS service. In the medium-power category, the primary purpose is point-to-point services, but space may be leased on these satellites for the provision of DBS services. In the low-power category, no official DBS services are provided. However, it was quickly discovered by home experimenters that a wide range of radio and TV programming could be received on this band, and it is now considered to provide a de facto DBS service, witness to which is the large number of *TV receive-only* (TVRO) dishes that have appeared in the yards and on the rooftops of homes in North America. TVRO reception of C-band signals in the home is prohibited in many other parts of the world, partly for aesthetic reasons, because of the comparatively large dishes used, and partly for commercial reasons. Many North American C-band TV broadcasts are now encrypted, or scrambled, to prevent unauthorized access, although this also seems to be spawning a new underground industry in descramblers. As shown in Table 1.4, true DBS service takes place in the Ku band. Figure 1.5 shows the components of a DBS system (Government of Canada, 1983). The television signal may be relayed over a terrestrial link to the uplink station. This transmits a very narrow beam signal to the satellite in the 14-GHz band. The satellite retransmits the television signal in a wide beam in the 12-GHz frequency band. Individual receivers within the beam coverage area will receive the satellite signal.

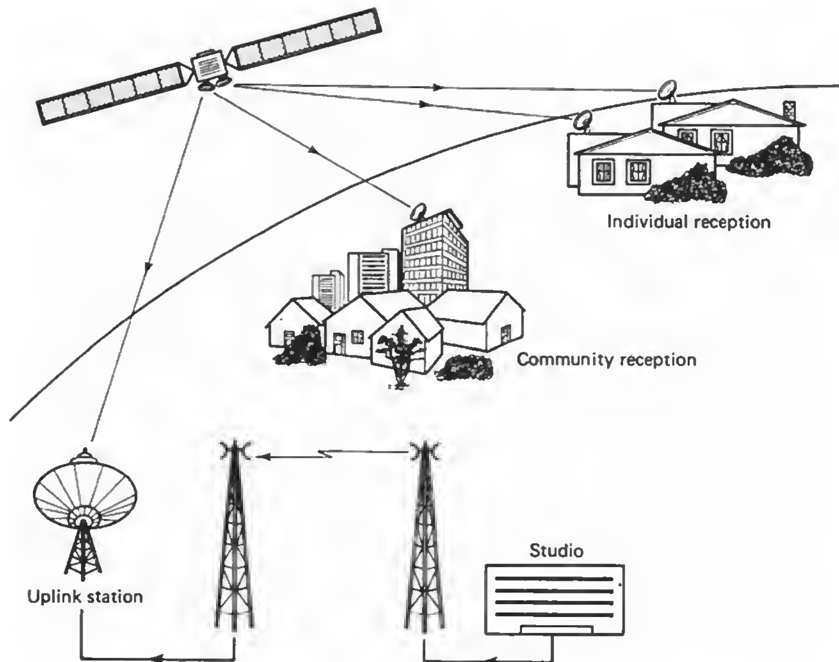


Figure 1.5 Components of a direct broadcasting satellite system. (*From Government of Canada, 1983, with permission.*)

Table 1.5 shows the orbital assignments for domestic fixed satellites for the United States (FCC, 1996). These satellites are in geostationary orbit, which is discussed further in Chap. 3. Table 1.6 shows the U.S. Ka-band assignments. Broadband services, such as Internet (see Chap. 15), can operate at Ka-band frequencies. In 1983, the U.S. FCC adopted a policy objective, setting 2° as the minimum orbital spacing for satellites operating in the 6/4-GHz band and 1.5° for those operating in the 14/12-GHz band (FCC, 1983). It is clear that interference between satellite circuits is likely to increase as satellites are positioned closer together. These spacings represent the minimum presently achievable in each band at acceptable interference levels. In fact, it seems likely that in some cases home satellite receivers in the 6/4-GHz band may be subject to excessive interference where 2° spacing is employed.

1.5 Polar Orbiting Satellites

Polar orbiting satellites orbit the earth in such a way as to cover the north and south polar regions. (Note that the term *polar orbiting* does not mean that the satellite orbits around one or the other of the poles).

TABLE 1.5 FCC Orbital Assignment Plan (May 7, 1996)

Location, degrees west longitude	Satellite	Band/polarization
139	Aurora Il/Satcom C-5	4/6 GHz (vertical)
139	ACS-3K (AMSC)	12/14 GHz
137	Satcom C-1	4/6 GHz (horizontal)
137	Unassigned	12/14 GHz
135	Satcom C-4	4/6 GHz (vertical)
135	Orion O-F4	12/14 GHz
133	Galaxy 1-R(S)	4/6 GHz (horizontal)
133	Unassigned	12/14 GHz
131	Satcom C-3	4/6 GHz (vertical)
131	Unassigned	12/14 GHz
129	Loral 1	4/6 GHz (horizontal)/12/14 GHz
127	Galaxy IX	4/6 GHz (vertical)
127	Unassigned	12/14 GHz
125	Galaxy 5-W	4/6 GHz (horizontal)
125	GSTAR II/unassigned	12/14 GHz
123	Galaxy X	4/6 GHz (vertical)/12/14 GHz
121	EchoStar FSS-2	12/14 GHz
105	GSTAR IV	12/14 GHz
103	GE-1	4/6 GHz (horizontal)
103	GSTAR 1/GE-1	12/14 GHz
101	Satcom SN-4 (formerly Spacenet IV-n)	4/6 GHz (vertical)/ 12/14 GHz
99	Galaxy IV(H)	4/6 GHz (horizontal)/12/14 GHz
97	Telstar 401	4/6 GHz (vertical)/12/14 GHz
95	Galaxy IlI(H)	4/6 GHz (horizontal)/12/14 GHz
93	Telstar 5	4/6 GHz (vertical)
93	GSTAR III/Telstar 5	12/14 GHz
91	Galaxy Vll(H)	4/6 GHz (horizontal)/12/14 GHz
89	Telestar 402R	4/6 GHz (vertical)/12/14 GHz
87	Satcom SN-3 (formerly Spacenet Ill-)/GE-4	4/6 GHz (horizontal)/12/14 GHz
85	Telstar 302/GE-2	4/6 GHz (vertical)
85	Satcom Ku-1/GE-2	12/14 GHz
83	Unassigned	4/6 GHz (horizontal)
83	EchoStar FSS-1	12/14 GHz
81	Unassigned	4/6 GHz (vertical)
81	Satcom Ku-2/unassigned	12/14 GHz
79	GE-5	4/6 GHz (horizontal)/12/14 GHz
77	Loral 2	4/6 GHz (vertical)/12/14 GHz
76	Comstar D-4	4/6 GHz (vertical)
74	Galaxy VI	4/6 GHz (horizontal)
74	SBS-6	12/14 GHz
72	Unassigned	4/6 GHz (vertical)
71	SBS-2	12/14 GHz
69	Satcom SN-2/Telstar 6	4/6 GHz (horizontal)/12/14 GHz
67	GE-3	4/6 GHz (vertical)/12/14 GHz
64	Unassigned	4/6 GHz (horizontal)
64	Unassigned	12/14 GHz
62	Unassigned	4/6 GHz (vertical)
62	ACS-2K (AMSC)	12/14 GHz
60	Unassigned	4/6 GHz
60	Unassigned	12/14 GHz

TABLE 1.6 Ka-Band Orbital Assignment Plan (FCC December 19, 1997)

Location	Company
147°W.L.	Morning Star Satellite Company, L.L.C.
125°W.L.	PanAmSat Licensee Corporation
121°W.L.	Echostar Satellite Corporation
115°W.L.	Loral Space & Communications, Ltd.
113°W.L.*	VisionStar, Inc.
109.2°W.L.	KaStar Satellite Communications Corp.
105°W.L.†	GE American Communications, Inc.
103°W.L.	PanAmSat Corporation
101°W.L.	Hughes Communications Galaxy, Inc.
99°W.L.	Hughes Communications Galaxy, Inc.
97°W.L.	Lockheed Martin Corporation
95°W.L.	NetSat 28 Company, L.L.C.
93°W.L.	Loral Space & Communications, Ltd.
91°W.L.	Comm, Inc.
89°W.L.	Orion Network Systems
87°W.L.	Comm, Inc.
85°W.L.	GE American Communications, Inc.
83°W.L.	Echostar Satellite Corporation
81°W.L.	Orion Network Systems
77°W.L.	Comm, Inc.
75°W.L.	Comm, Inc.
73°W.L.	KaStar Satellite Corporation
67°W.L.	[under consideration]
62°W.L.	Morning Star Satellite Company, L.L.C.
58°W.L.	PanAmSat Corporation
49°W.L.	Hughes Communications Galaxy, Inc.
47°W.L.	Orion Atlantic, L.P.
21.5°W.L.	Lockheed Martin Corporation
17°W.L.	GE American Communications, Inc.
2°E.L.	Lockheed Martin Corporation
25°E.L.	Hughes Communication Galaxy, Inc.
30°E.L.	Morning Star Satellite Company, L.L.C.
36°E.L.	PanAmSat Corporation
40°E.L.	PanAmSat Corporation
48°E.L.	PanAmSat Corporation
54°E.L.	Hughes Communications Galaxy, Inc.
56°E.L.	GE American Communications, Inc.
78°E.L.	Orion Network Systems, Inc.
101°E.L.	Hughes Communications Galaxy, Inc.
105.5°E.L.	Loral Space & Communications, Ltd.
107.5°E.L.	Morning Star Satellite Company, L.L.C.

NOTES: FCC—Federal Communications Commission; W.L.—West Longitude; E.L.—East Longitude.

*VisionStar will operate at a nominal orbit location of 113.05°W.L., and with a station-keeping box of ±0.05°, thereby increasing the separation from the Canadian filing at 111.1°W.L. by 0.1° relative to the worst case orbital spacing using the station keeping assumed in the ITU publications (±0.1°).

†The applicants in the range 95° to 105°W.L. have agreed to operate their satellites with a nominal 0.05° offset to the west, and with a station-keeping box of ±0.05°, thereby increasing the separation from the Luxembourg satellite at 93.2°W.L. by 0.1° relative to the worst case orbital spacing using the station keeping assumed in the ITU publications (±0.1°).

TABLE 1.6 Ka-Band Orbital Assignment Plan (FCC December 19, 1997) (*Continued*)

Location	Company
111°E.L.	Hughes Communications Galaxy, Inc.
114.5°E.L.	GE American Communications, Inc.
124.5°E.L	PanAmSat Corporation
126.5°E.L.	Orion Network Systems, Inc.
130°E.L.	Lockheed Martin Corporation
149°E.L.	Hughes Communications Galaxy, Inc.
164°E.L.	Hughes Communications Galaxy, Inc.
173°E.L.	PanAmSat Corporation
175.25°E.L.	Lockheed Martin Corporation

Figure 1.6 shows a polar orbit in relation to the geostationary orbit. Whereas there is only one geostationary orbit, there are, in theory, an infinite number of polar orbits. The U.S. experience with weather satellites has led to the use of relatively low orbits, ranging in altitude between 800 and 900 km, compared with 36,000 km for the geostationary orbit. *Low earth orbiting* (LEO) satellites are known generally by the acronym LEOSATS.

In the United States, the *National Polar-orbiting Operational Environmental Satellite System* (NPOESS) was established in 1994 to consolidate the polar satellite operations of the Air Force, NASA (*National Aeronautics and Space Administration*) and NOAA (*National Oceanic and Atmospheric Administration*). NPOESS manages the

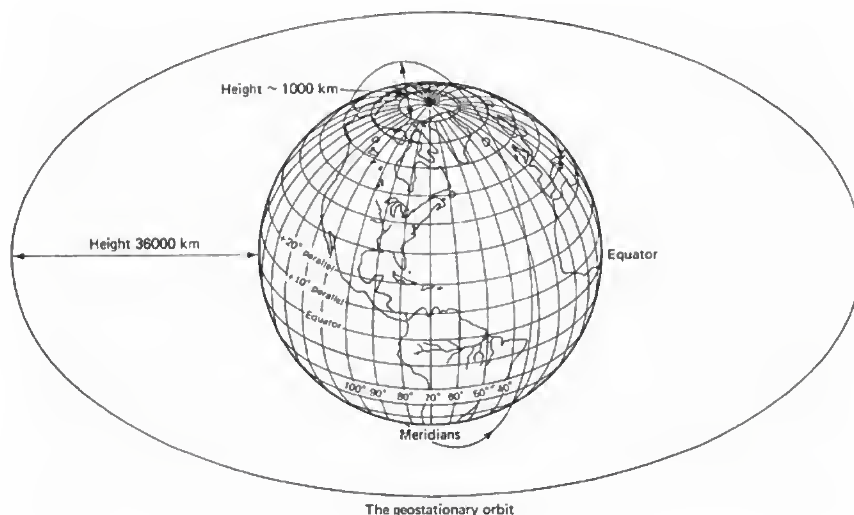


Figure 1.6 Geostationary orbit and one possible polar orbit.

Integrated Program Office (IPO) and the Web page can be found at *http://www.ipo.noaa.gov/*. As of 2005, a four-orbit system is in place, consisting of two U.S. Military orbits, one U.S. Civilian orbit and one EUMETSAT/METOP orbit. Here, METSAT stands for *meteorological satellite* and EUMETSAT stands for the *European organization for the exploration of the METSAT program*. METOP stands for *meteorological operations*. These orbits are sun synchronous, meaning that they cross the equator at the same local time each day. For example, the satellites in the NPOESS (civilian) orbit will cross the equator, going from south to north, at times 1:30 P.M., 5:30 P.M., and 9:30 P.M. Sun-synchronous orbits are described in more detail in Chap. 2, but briefly, the orbit is arranged to rotate eastward at a rate of 0.9856°/day, to make it *sun synchronous*. In a sun-synchronous orbit the satellite crosses the same spot on the earth at the same local time each day, so that the same area of the earth can be viewed under approximately the same lighting conditions each day. A sun-synchronous orbit is inclined slightly to the west of the north pole. By definition, an orbital pass from south to north is referred to as an *ascending pass*, and from north to south as a *descending pass*.

The polar orbits are almost circular, and as previously mentioned they are at a height of between 800 and 900 km above earth. The polar orbiters are able to track weather conditions over the entire earth, and provide a wide range of data, including visible and infrared radiometer data for imaging purposes, radiation measurements, and temperature profiles. They carry ultraviolet sensors that measure ozone levels, and they can monitor the ozone hole over Antarctica. The polar orbiters carry a NOAA letter designation before launch, which is changed to a numeric designation once the satellite achieves orbit. For example, NOAA M, launched on June 24, 2002, became NOAA 17 when successfully placed in orbit. The series referred to as the *KLM satellites* carry much improved instrumentation. Some details are shown in Table 1.7.

Most of the polar orbiting satellites used in weather and environmental studies, and as used for monitoring and in search and rescue, have a "footprint" about 6000 km in diameter. This is the size of the antenna spot beam on the surface of the earth. As the satellite orbits the earth, the spot beam sweeps out a swath on the earth's surface about 6000 km wide passing over north and south poles. The orbital period of these satellites is about 102 min. Since a day has 1440 min, the number of orbits per day is 1440/102 or approximately 14. In the 102 min the earth rotates eastward 360° × 102/1440 or about 25°. Neglecting for the moment the small eastward rotation of the orbit required for sun synchronicity, the earth will rotate under the subsatellite path by this amount, as illustrated in Fig. 1.7.

TABLE 1.7 NOAA KLM Satellites

Launch date	NOAA-K (NOAA-15): May 13, 1998 NOAA-L: September 21, 2000 NOAA-M: June 24, 2000 NOAA-N: March 19, 2005 (tentative) NOAA-N': July 2007
Mission life	2 years minimum
Orbit	Sun synchronous, 833 ± 19 km or 870 ± 19 km
Mass	1478.9 kg on orbit; 2231.7 kg on launch
Length/Diameter	4.18 m/1.88 m
Sensors	Advanced very high resolution radiometer (AVHRR/3) Advanced microwave sounding unit-A (AMSU-A) Advanced microwave sounding unit-B (AMSU-B) High resolution infrared radiation sounder (HIRS/3) Space environment monitor (SEM/2) Search and rescue (SAR) repeater and processor Data collection system (DCS/2)

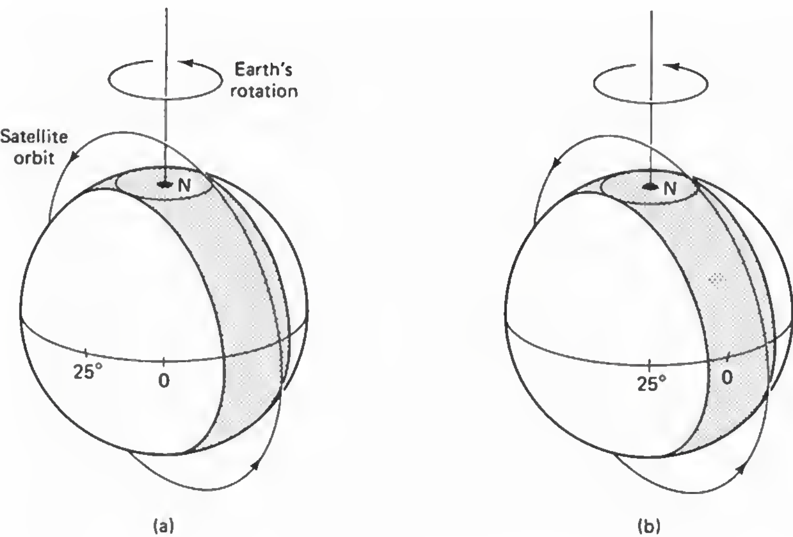


Figure 1.7 Polar orbiting satellite: (*a*) first pass; (*b*) second pass, earth having rotated 25°. Satellite period is 102 min.

1.6 Argos System

The Argos *data collection system* (DCS) collects environmental data radioed up from *platform transmitter terminals* (PTT) (Argos, 2005). The characteristics of the PTT are shown in Table 1.8.

The transmitters can be installed on many kinds of platforms, including fixed and drifting buoys, balloons, and animals. The physical size of the transmitters depends on the application. These can weigh as little as 17 g for transmitters fitted to birds, to track their migratory patterns. The PTTs transmit automatically at preset intervals, and those within the 6000 km swath are received by the satellite. As mentioned, the satellite completes about 14 orbits daily, and all orbits cross over the poles. A PTT located at the polar regions would therefore be able to deliver approximately 14 messages daily. At least two satellites are operational at any time, which doubles this number to 28. At the equator the situation is different. The equatorial radius of the earth is approximately 6378 km, which gives a circumference of about 40,074 km. Relative to the orbital footprint, a given longitude at the equator will therefore rotate with the earth a distance of $40074 \times 102/1440$ or about 2839 km. This assumes a stationary orbital path, but as mentioned previously the orbit is sun synchronous, which means that it rotates eastward almost 1° per day (see Sec. 2.8.1), that is in the same direction as the earth's rotation. The overall result is that an equatorial PTT starting at the western edge of the footprint swath will "see" between three and four passes per day for one satellite. Hence the equatorial passes number between six and seven per day for two satellites. During any one pass the PTT is in contact with the satellite for 10 min on average. The messages received at the satellite are retransmitted in "real time" to one of a number of regional ground receiving stations whenever the satellite is within range. The messages are also stored aboard the satellites on tape recorders, and are "dumped" to one of three main ground receiving stations. These are located at Wallops Island, VA, USA, Fairbanks, Alaska, USA, and Lannion, France. The Doppler shift in the frequency received at the satellite is used to determine the location of the PTT. This is discussed further in connection with the Cospas-Sarsat search and rescue satellites.

TABLE 1.8 Platform Transmitter Terminals (PTT) Characteristics

Uplink frequency	401.75 MHz
Message length	Up to 32 bytes
Repetition period	45–200 s
Messages/pass	Varies depending on latitude and type of service
Transmission time	360–920 ms
Duty cycle	Varies
Power	Battery, solar, external

SOURCE: www.argosinc.com/documents/sysdesc.pdf

1.7 Cospas-Sarsat*

COSPAS is an acronym from the Russian *Cosmicheskaya Sistyema Poiska Avariynich Sudov*, meaning space system for the search of vessels in distress and SARSAT stands for *Search and Rescue Satellite-Aided Tracking* (see *http://www.equipped.com/cospas-sarsat_overview.htm*). The initial Memorandum of Understanding that led to the development of the system was signed in 1979 by agencies from Canada, France, the USA, and the former USSR. There are (as of November 2004) 37 countries and organizations associated with the program. Canada, France, Russia and the USA provide and operate the satellites and ground-segment equipment, and other countries provide ground-segment support. A full list of participating countries will be found in Cospas-Sarsat (2004). The system has now been developed to the stage where both low earth orbiting (LEO) satellites and *geostationary earth orbiting* (GEO) satellites are used, as shown in Fig. 1.8.

The basic system requires users to carry distress radio beacons, which transmit a carrier signal when activated. A number of different beacons are available: *emergency locator transmitter* (ELT) for aviation use; *emergency position indicating radio beacon* (EPIRB) for maritime use; and *personal locator beacon* (PLB) for personal use. The beacons can be activated manually or automatically (e.g., by a crash sensor). The transmitted signal is picked up by a LEO satellite, and because this satellite is moving relative to the radio beacon, a *Doppler shift* in frequency is observed. In effect, if the line of sight distance between transmitter and satellite is shortened as a result of the relative motion, the wavelength of the emitted signal is also shortened. This in turn means the received frequency is increased. If the line of sight distance is lengthened as a result of the

*http://www.cospas-sarsat.org/

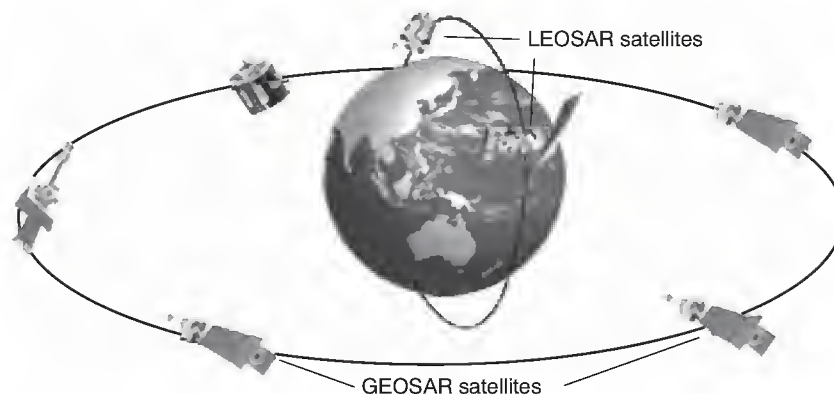


Figure 1.8 Geostationary orbit search and rescue (GEOSAR) and low earth orbit search and rescue (LEOSAR) satellites. (*Courtesy of Cospas-Sarsat Secretariat.*)

relative motion the wavelength is lengthened and therefore the received frequency decreased. It should be kept in mind that the radio-beacon emits a constant frequency, and the electromagnetic wave travels at constant velocity, that of light. Denoting the constant emitted frequency by f_0, the relative velocity between satellite and beacon, measured along the line of sight as v, and the velocity of light as c, then to a close approximation the received frequency is given by (assuming $v<<c$):

$$f = \left(1 + \frac{v}{c}\right)f_0 \tag{1.1}$$

The relative velocity v is positive when the line of sight distance is decreasing, (satellite and beacon moving closer together) and negative when it is increasing (satellite and beacon moving apart). The relative velocity v is a function of the satellite motion and of the earth's rotation. The frequency difference resulting from the relative motion is

$$\Delta f = f - f_0 = \frac{v}{c} f_0 \tag{1.2}$$

The fractional change is

$$\frac{\Delta f}{f_0} = \frac{v}{c} \tag{1.3}$$

When v is zero, the received frequency is the same as the transmitted frequency. When the beacon and satellite are approaching each other, v is positive, which results in a positive value of Δf. When the beacon and satellite are receding, v is negative, resulting in a negative value of Δf. The time at which Δf is zero is known as the *time of closest approach*.

Figure 1.9 shows how the beacon frequency, as received at the satellite, varies for different passes. In all cases, the received frequency goes from being higher to being lower than the transmitted value as the satellite approaches and then recedes from the beacon. The longest record and the greatest change in frequency are obtained if the satellite passes over the site, as shown for pass no. 2. This is so because the satellite is visible for the longest period during this pass. Knowing the orbital parameters for the satellite, the beacon frequency, and the Doppler shift for any one pass, the distance of the beacon relative to the projection of the orbit on the earth can be determined. However, whether the beacon is east or west of the orbit cannot be determined easily from a single pass. For two successive passes, the effect of the earth's rotation on the Doppler shift can be estimated more accurately, and from this it can be determined whether the orbital path is moving closer to, or moving away from the beacon. In this way, the ambiguity in east-west positioning is resolved. The satellite must of course get the information back to an earth station so that the search and rescue operation can be completed, successfully one hopes. The SARSAT communicates on a

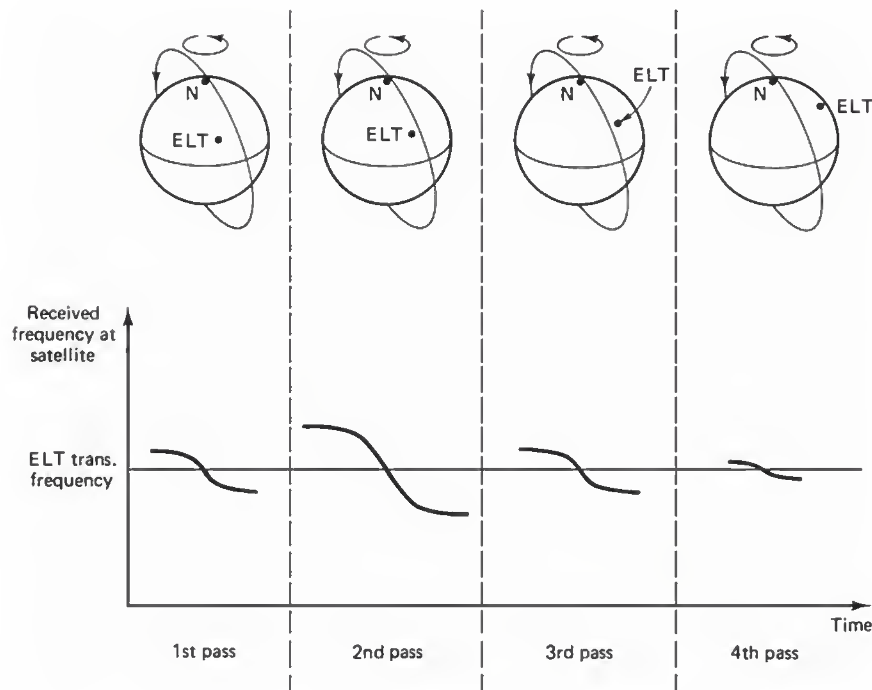


Figure 1.9 Showing the Doppler shift in received frequency on successive passes of the satellite. ELT—emergency locator transmitter.

downlink frequency of 1544.5 MHz to one of several *local user terminals* (LUTs) established at various locations throughout the world.

In the original Cospas-Sarsat system, the signal from the emergency radio beacons was at a frequency of 121.5 MHz. It was found that over 98 percent of the alerts at this frequency were false, often being caused by interfering signals from other services and by inappropriate handling of the equipment. The 121.5-MHz system relies entirely on the Doppler shift, and the carrier does not carry any identification information. The power is low, typically a few tenths of a watt, which limits locational accuracy to about 10 to 20 km. There are no signal storage facilities aboard the satellites for the 121.5-MHz signals, which therefore requires that the distress site (the distress beacon) and the LUT must be visible simultaneously from the satellite. Because of these limitations, the 121.5-MHz beacons are being phased out, and the 121.5-MHz service will terminate on February 1, 2009. Cospas-13, planned for launch in 2006, and Sarsat-14, planned for launch from 2009, will not carry 121.5-MHz beacons. However, all Cospas-Sarsat satellites launched prior to these will carry the 121.5-MHz processors. (Recall that Sarsat-7 is NOAA-15, Sarsat-8 is NOAA-L, Sarsat-9 is NOAA-M, and Sarsat-10 is NOAA-N.)

Newer beacons operating at a frequency of 406 MHz are being introduced. The power has been increased to 5 W, which should permit locational accuracy to 3 to 5 km (Scales and Swanson, 1984). The 406-MHz carrier is modulated with information such as an identifying code, the last known position, and the nature of the emergency. The satellite has the equipment for storing and forwarding the information from a continuous memory dump, providing complete worldwide coverage with 100 percent availability. The polar orbiters, however, do not provide continuous coverage. The mean time between a distress alert being sent and the appropriate search and rescue coordination center being notified is estimated at 27 min satellite storage time plus 44 min waiting time for a total delay of 71 min (Cospas-Sarsat, 1994a, b).

The nominal frequency is 406 MHz, and originally, a frequency of 406.025 MHz was used. Because of potential conflict with the GEOSAR system, new channels at 406.028 MHz and 406.037 MHz have been opened, and type approval for the 406.025 MHz channel ceased in January 2002. However, beacon types approved before the January 2001 date and still in production may continue to operate at 406.025 MHz. More details will be found at http://www.cospas-sarsat.org/Beacons/406Bcns.htm.

The status of the Cospas-Sarsat *low earth orbiting search and rescue* (LEOSAR) satellites as of April 19, 2005 is shown in Table 1.9.

TABLE 1.9 Status of LEOSAR Payload Instruments

Satellite	Repeater instruments, MHz			Search and rescue processor		Comments
	121.5	243	406	Global	Local	
Sarsat 6	F	F	F	NO	NO	
Sarsat 7	F	L	F	F	F	Intermittent loss of the 243-MHz service, which may affect an entire or partial satellite pass.
Sarsat 8	L	NO	F	F	F	Intermittent loss of the 121.5-MHz service, which may affect an entire or partial satellite pass.
Sarsat 9	F	F	F	F	F	
Cospas 4	NO	NA		NO	NO	
Cospas 5	F	NA		NO	NO	

NOTES: F—fully operational; L—limited operation; NO—not operational; NA—not applicable.
SOURCE: http://www.cospas-sarsat.org/Status/spaceSegmentStatus.htm

The nominal space segment of LEOSAR consists of four satellites, although as shown in Table 1.9 more satellites may be in service at any one time. The status of the 121.5-MHz LEOSAR system as of November 2004 consisted of repeaters on five polar orbiters, 43 ground receiving stations (referred to as *LEOSAR local user terminals,* or LEOLUTs), 26 *mission control centers* (MCCs), and about 680,000 beacons operating at 121.5 MHz, carried mostly on aircraft and small vessels. The MCC alerts the *rescue coordination center* (RCC) nearest the location where the distress signal originated, and the RCC takes the appropriate action to effect a rescue. In 2006, Cospas-11 will be launched aboard a new LEO satellite called "Sterkh." This is a small, 190 kg satellite designed specifically for Cospas-Sarsat operations.

The status of the GEOSAR segment of the Cospas-Sarsat system is shown in Table 1.10.

Since the geostationary satellites are by definition stationary with respect to the earth, there is no Doppler shift of the received beacon carrier. The 406-MHz beacons for the GEOSAR component carry positional information obtained from the global navigational satellite systems such as the American GPS (see Sec. 17.5) system, the Russian *global navigation satellite system* (GLONASS) and Galileo (European). These navigational systems employ *medium earth orbiting* (MEO) satellites, and the space agencies responsible for these navigational systems have plans to include 406-MHz repeaters on the MEO satellites.

Although the GEOSAR system provides wide area coverage it does not cover the polar regions, the antenna "footprint" being limited to latitudes of about 75° N and S. The coverage areas are shown in Fig. 1.10.

TABLE 1.10 Status of GEOSAR Payload Instruments

Satellite	Status	Gain control	Comments
GOES-9 (155°E)	F	Fixed	
GOES-East (75°W)	F	AGC	
GOES-West (135°W)	F	Fixed	
INSAT 3A (93.5°E)	L	TBD	INSAT 3A is currently under test; however, alerts from the system are distributed operationally. INSAT system does not process second protected field of long format messages.
MSG-1 (3.4°W)	F	Fixed	

NOTES: F—fully operational; L—limited operation; NO—not operational; NA—not applicable; TBD—to be determined; GOES—Geostationary Operational Environmental Satellite (USA); INSAT—Indian Satellite; MSG—Meteosat Second Generation (European).
SOURCE: http: //www.cospas-sarsat.org/Status/spaceSegmentStatus.htm

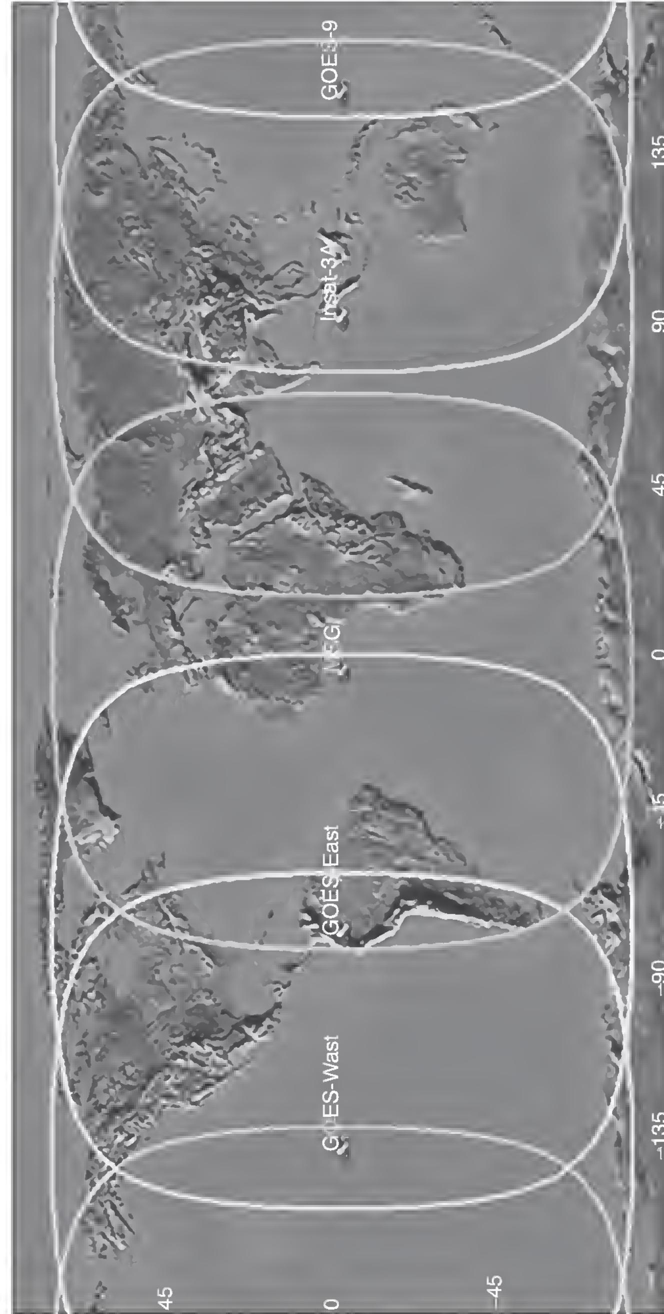


Figure 1.10 GEOSAR coverage.

1.8 Problems

1.1. Describe briefly the main advantages offered by satellite communications. Explain what is meant by a *distance-insensitive communications system*.

1.2. Comparisons are sometimes made between satellite and optical fiber communications systems. State briefly the areas of application for which you feel each system is best suited.

1.3. Describe briefly the development of INTELSAT starting from the 1960s through the present. Information can be found at Web site http://www.intelsat.com/.

1.4. From the Web site http://www.intelsat.com/, find the positions of the INTELSAT 901 and the INTELSAT 10-02 satellites, as well as the number of C-band and Ku-band transponders on each.

1.5. From Table 1.3, and by accessing the Intelsat web site, determine which satellites provide service to each of the regions AOR, IOR, and POR.

1.6. Referring to Table 1.4, determine the power levels, in watts, for each of the three categories listed.

1.7. From Table 1.5, determine typical orbital spacing in degrees for (*a*) the 6/4-GHz band and (*b*) the 14/12-GHz band.

1.8. Give reasons why the Ku band is used for the DBS service.

1.9. An earth station is situated at longitude 91°W and latitude 45°N. Determine the range to the Galaxy VII satellite. A spherical earth of uniform mass and mean radius 6371 km may be assumed.

1.10. Given that the earth's equatorial radius is 6378 km and the height of the geostationary orbit is 36,000 km, determine the intersatellite distance between the VisionStar Inc. satellite and the NetSat 28 Company L.L.C. satellite, operating in the Ka band.

1.11. Explain what is meant by a *polar orbiting satellite*. A NOAA polar orbiting satellite completes one revolution around the earth in 102 min. The satellite makes a north to south equatorial crossing at longitude 90°W. Assuming that the orbit is circular and crosses exactly over the poles, estimate the position of the subsatellite point at the following times after the equatorial crossing: (*a*) 0 h, 10 min; (*b*) 1 h, 42 min; (*c*) 2 h, 0 min. A spherical earth of uniform mass may be assumed.

1.12. By accessing the NOAA Web page at http://www.noaa.gov/, find out how the GOES take part in weather forecasting. Give details of the GOES-12 characteristics.

1.13. The Cospas-Sarsat Web site is at http://www.cospas-sarsat.org. Access this site and find out the number and location of the LEOLUTs in current use.

1.14. Using information obtained from the Cospas-Sarsat Web site, find out which satellites carry (*a*) 406-MHz *SAR processors* (SARPs), (*b*) 406-MHz *SAR repeaters* (SARRs), and (*c*) 121.5-MHz SARRs. What is the basic difference between a SARP and a SARR?

1.15. Intelsat satellite 904 is situated at 60°E. Determine the land areas (markets)the satellite can service. The global EIRP is given as 31.0 up to 35.9 dBW, beam edge to beam peak. What are the equivalent values in watts? (see App. G for the definition of dBW).

1.16. A satellite is in a circular polar orbit at a height of 870 km, the orbital period being approximately 102 min. Assuming an average value of earth's radius of 6371 km determine approximately the maximum period the satellite is visible from a beacon at sea level.

1.17. A satellite is in a circular polar orbit at a height of 870 km, the orbital period being approximately 102 min. The satellite orbit passes directly over a beacon at sea level. Assuming an average value of earth's radius of 6371 km determine approximately the fractional Doppler shift at the instant the satellite is first visible from the beacon.

References

Argos. 2005. General information to info@argosinc.com Customer support to DUS, at www.argosinc.com/documents. From the menu list select sysdesc.pdf

Brown, M. P., Jr. (Ed.). 1981. *Compendium of Communication and Broadcast Satellites 1958 to 1980*. IEEE Press, New York.

Cospas-Sarsat, at http://www.cospas-sarsat.org/

Cospas-Sarsat, at http://www.cospas-sarsat.org/Beacons/406Bcns.htm

Cospas-Sarsat, at http://www.cospas-sarsat.org/Status/spaceSegmentStatus.htm

Cospas-Sarsat, at http://www.equipped.com/cospas-sarsat_overview.htm

Cospas-Sarsat. 1994a. *System Data No. 17*, February.

Cospas-Sarsat. 1994b. *Information Bulletin No. 8*, February.

Cospas-Sarsat. 2004. *Information Bulletin No.17*, November, at http://www.cospas-sarsat.org

FCC. 1983. "Licensing of Space Stations in the Domestic Fixed-Satellite Service and Related Revisions of Part 25 of the Rules and Regulations, Report 83-184 33206, CC Docket 81-184." Federal Communications Commission, Washington, DC.

FCC. 1997 "Assignment of Orbital Locations to Space Stations in the Ka-Band." Adopted May 8, 1997, released May 9, at http://www.fcc.gov/Bureaus/International/Orders/1997/da970967.txt

FCC. 1996. "Orbital Assignment Plan," at http://www.fcc.gov/Bureaus/International/Orders/1996/da960713.txt

Government of Canada. 1983. "Direct-to-Home Satellite Broadcasting for Canada." *IEEE Spectrum*, March.

Information Services, Department of Communications. Intelsat, at http://www.intelsat.com/

Lilly, C. J. 1990. "INTELSAT's New Generation." *IEE Review*, Vol. 36, No. 3, March.

NOAA, at http://www.ipo.noaa.gov/

Pritchard, W. L. 1984. "The History and Future of Commercial Satellite Communications." *IEEE Commun. Mag.,* Vol. 22, No. 5, May, pp. 22–37.

Reinhart, E. E. 1990. "Satellite Broadcasting and Distribution in the United States." *Telecommun. J.,* Vol. 57, No. V1, June, pp. 407–418.

Sachdev, D. K., P. Nadkarni, P. Neyret, L. R. Dest, K. Betaharon, and W. J. English. 1990. "INTELSAT V11: A Flexible Spacecraft for the 1990s and Beyond." *Proc. IEEE,* Vol. 78, No. 7, July, pp. 1057–1074.

Scales, W. C., and R. Swanson. 1984. "Air and Sea Rescue via Satellite Systems." IEEE Spectrum, March, pp. 48–52.

Chapter

2

Orbits and Launching Methods

2.1 Introduction

Satellites (spacecraft) orbiting the earth follow the same laws that govern the motion of the planets around the sun. From early times much has been learned about planetary motion through careful observations. Johannes Kepler (1571–1630) was able to derive empirically three laws describing planetary motion. Later, in 1665, Sir Isaac Newton (1642–1727) derived Kepler's laws from his own laws of mechanics and developed the theory of gravitation [for very readable accounts of much of the work of these two great men, see Arons (1965) and Bate et al. (1971)].

Kepler's laws apply quite generally to any two bodies in space which interact through gravitation. The more massive of the two bodies is referred to as the *primary,* the other, the *secondary* or *satellite.*

2.2 Kepler's First Law

Kepler's first law states that the path followed by a satellite around the primary will be an ellipse. An ellipse has two focal points shown as F_1 and F_2 in Fig. 2.1. The center of mass of the two-body system, termed the *barycenter,* is always centered on one of the foci. In our specific case, because of the enormous difference between the masses of the earth and the satellite, the center of mass coincides with the center of the earth, which is therefore always at one of the foci.

The semimajor axis of the ellipse is denoted by a, and the semiminor axis, by b. The eccentricity e is given by

$$e = \frac{\sqrt{a^2 - b^2}}{a} \tag{2.1}$$

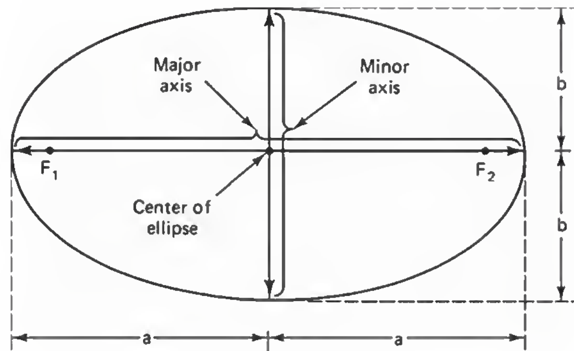


Figure 2.1 The foci F_1 and F_2, the semimajor axis a, and the semiminor axis b of an ellipse.

The eccentricity and the semimajor axis are two of the orbital parameters specified for satellites (spacecraft) orbiting the earth. For an elliptical orbit, $0 < e < 1$. When $e = 0$, the orbit becomes circular. The geometrical significance of eccentricity, along with some of the other geometrical properties of the ellipse, is developed in App. B.

2.3 Kepler's Second Law

Kepler's second law states that, for equal time intervals, a satellite will sweep out equal areas in its orbital plane, focused at the barycenter. Referring to Fig. 2.2, assuming the satellite travels distances S_1 and S_2 meters in 1 s, then the areas A_1 and A_2 will be equal. The average velocity in each case is S_1 and S_2 m/s, and because of the equal area law, it follows that the velocity at S_2 is less than that at S_1. An important

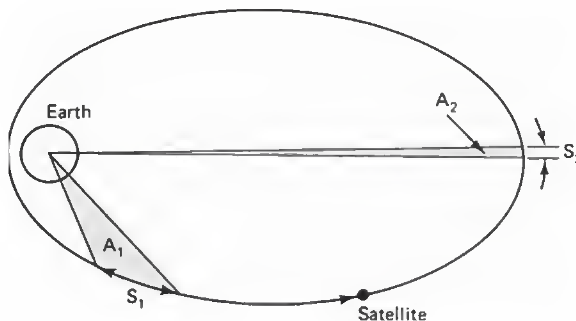


Figure 2.2 Kepler's second law. The areas A_1 and A_2 swept out in unit time are equal.

consequence of this is that the satellite takes longer to travel a given distance when it is farther away from earth. Use is made of this property to increase the length of time a satellite can be seen from particular geographic regions of the earth.

2.4 Kepler's Third Law

Kepler's third law states that the square of the periodic time of orbit is proportional to the cube of the mean distance between the two bodies. The mean distance is equal to the semimajor axis a. For the artificial satellites orbiting the earth, Kepler's third law can be written in the form

$$a^3 = \frac{\mu}{n^2} \tag{2.2}$$

where n is the mean motion of the satellite in radians per second and μ is the earth's geocentric gravitational constant. Its value is (see Wertz, 1984, Table L3).

$$\mu = 3.986005 \times 10^{14} \text{ m}^3/\text{s}^2 \tag{2.3}$$

Equation (2.2) applies only to the ideal situation of a satellite orbiting a perfectly spherical earth of uniform mass, with no perturbing forces acting, such as atmospheric drag. Later, in Sec. 2.8, the effects of the earth's oblateness and atmospheric drag will be taken into account.

With n in radians per second, the orbital period in seconds is given by

$$P = \frac{2\pi}{n} \tag{2.4}$$

The importance of Kepler's third law is that it shows there is a fixed relationship between period and semimajor axis. One very important orbit in particular, known as the *geostationary orbit,* is determined by the rotational period of the earth and is described in Chap. 3. In anticipation of this, the approximate radius of the geostationary orbit is determined in the following example.

Example 2.1 Calculate the radius of a circular orbit for which the period is 1 day.

Solution There are 86,400 seconds in 1 day, and therefore the mean motion is

$$\begin{aligned} n &= \frac{2\pi}{86400} \\ &= 7.272 \times 10^{-5} \text{ rad/s} \end{aligned}$$

From Kepler's third law:

$$a = \left[\frac{3.986005 \times 10^{14}}{(7.272 \times 10^{-5})^2}\right]^{1/3}$$

$$= \underline{\underline{42{,}241 \text{ km}}}$$

Since the orbit is circular the semimajor axis is also the radius.

2.5 Definitions of Terms for Earth-Orbiting Satellites

As mentioned previously, Kepler's laws apply in general to satellite motion around a primary body. For the particular case of earth-orbiting satellites, certain terms are used to describe the position of the orbit with respect to the earth.

Subsatellite path. This is the path traced out on the earth's surface directly below the satellite.

Apogee. The point farthest from earth. Apogee height is shown as h_a in Fig. 2.3.

Perigee. The point of closest approach to earth. The perigee height is shown as h_p in Fig. 2.3.

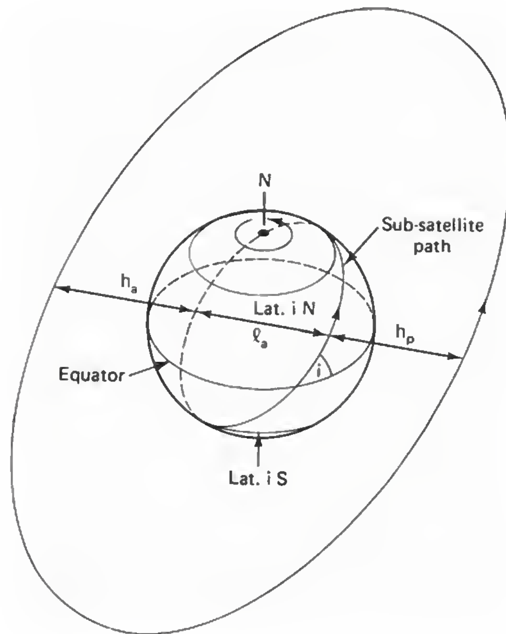


Figure 2.3 Apogee height h_a, perigee height h_p, and inclination i. l_a is the line of apsides.

Line of apsides. The line joining the perigee and apogee through the center of the earth.

Ascending node. The point where the orbit crosses the equatorial plane going from south to north.

Descending node. The point where the orbit crosses the equatorial plane going from north to south.

Line of nodes. The line joining the ascending and descending nodes through the center of the earth.

Inclination. The angle between the orbital plane and the earth's equatorial plane. It is measured at the ascending node from the equator to the orbit, going from east to north. The inclination is shown as i in Fig. 2.3. It will be seen that the greatest latitude, north or south, reached by the subsatellite path is equal to the inclination.

Prograde orbit. An orbit in which the satellite moves in the same direction as the earth's rotation, as shown in Fig. 2.4. The prograde orbit is also known as a *direct orbit.* The inclination of a prograde orbit always lies between 0° and 90°. Most satellites are launched in a prograde orbit because the earth's rotational velocity provides part of the orbital velocity with a consequent saving in launch energy.

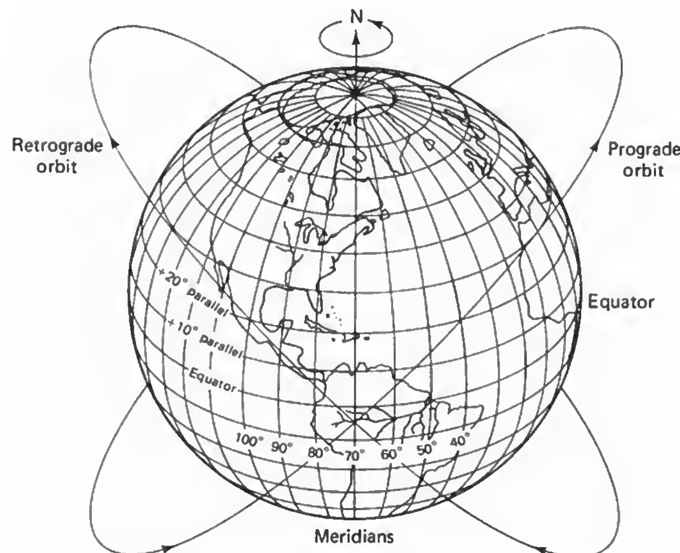


Figure 2.4 Prograde and retrograde orbits.

Retrograde orbit. An orbit in which the satellite moves in a direction counter to the earth's rotation, as shown in Fig. 2.4. The inclination of a retrograde orbit always lies between 90° and 180°.

Argument of perigee. The angle from ascending node to perigee, measured in the orbital plane at the earth's center, in the direction of satellite motion. The argument of perigee is shown as ω in Fig. 2.5.

Right ascension of the ascending node. To define completely the position of the orbit in space, the position of the ascending node is specified. However, because the earth spins, while the orbital plane remains stationary (slow drifts that do occur are discussed later), the longitude of the ascending node is not fixed, and it cannot be used as an absolute reference. For the practical determination of an orbit, the longitude and time of crossing of the ascending node are frequently used. However, for an absolute measurement, a fixed reference in space is required. The reference chosen is the *first point of Aries,* otherwise known as the vernal, or spring, equinox. The vernal equinox occurs when the sun crosses the equator going from south to north, and an imaginary line drawn from this equatorial crossing through the center of the sun points to the first point of Aries (symbol ♈). This is the *line of Aries.* The right ascension of the ascending node is then the angle measured eastward, in the equatorial plane, from the ♈ line to the ascending node, shown as Ω in Fig. 2.5.

Mean anomaly. Mean anomaly M gives an average value of the angular position of the satellite with reference to the perigee. For a

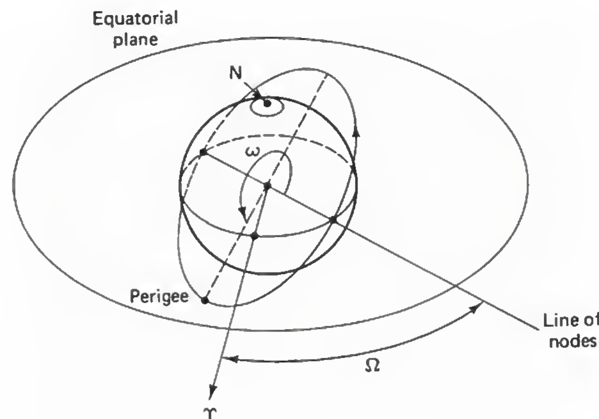


Figure 2.5 The argument of perigee ω and the right ascension of the ascending node Ω.

circular orbit, M gives the angular position of the satellite in the orbit. For elliptical orbit, the position is much more difficult to calculate, and M is used as an intermediate step in the calculation as described in Sec. 2.9.5.

True anomaly. The true anomaly is the angle from perigee to the satellite position, measured at the earth's center. This gives the true angular position of the satellite in the orbit as a function of time. A method of determining the true anomaly is described in Sec. 2.9.5.

2.6 Orbital Elements

Earth-orbiting artificial satellites are defined by six orbital elements referred to as the *keplerian element set.* Two of these, the semimajor axis a and the eccentricity e described in Sec. 2.2, give the shape of the ellipse. A third, the mean anomaly M_0, gives the position of the satellite in its orbit at a reference time known as the *epoch.* A fourth, the argument of perigee ω, gives the rotation of the orbit's perigee point relative to the orbit's line of nodes in the earth's equatorial plane. The remaining two elements, the inclination i and the right ascension of the ascending node Ω, relate the orbital plane's position to the earth. These four elements are described in Sec. 2.5.

Because the equatorial bulge causes slow variations in ω and Ω, and because other perturbing forces may alter the orbital elements slightly, the values are specified for the reference time or epoch, and thus the epoch also must be specified.

Appendix C lists the two-line elements provided to users by the U.S. *National Aeronautics and Space Administration* (NASA). The two-line elements may be downloaded from Celestrak at http://celestrak.com/NORAD/elements/. Figure 2.6 shows how to interpret the NASA two-line elements.

It will be seen that the semimajor axis is not specified, but this can be calculated from the data given. An example calculation is presented in Example 2.2.

Example 2.2 Calculate the semimajor axis for the satellite parameters given in Table 2.1.

Solution The mean motion is given in Table 2.1 as NN = 14.23304826 day^{-1}
In rad/s this is

$$
\begin{aligned}
n_0 &= 2 \times \pi \times \text{NN} \\
&= 0.00104\ \text{s}^{-1}
\end{aligned}
$$

year
day
dn/dt

```
1  25338U  98030A   00223.79688452  .00000307  00000-0  15447-3 0  8787
2  25338  98.6328 251.5324 0011501 113.5534 246.6853 14.23304826116635
```

3 7 9 16 18 19 21 25 27 32 33 35 42 44 51 53 63

i
Ω
e
ω
M
n
Rev. No.

Figure 2.6 Two-line elements for NOAA-15.

TABLE 2.1 Details from the NASA Bulletins (see Fig. 2.6 and App. C)

Line no.	Columns	Description
1	3–7	*Satellite number*: 25338
1	19–20	*Epoch year* (last two digits of the year): 00
1	21–32	*Epoch day* (day and fractional day of the year): 223.79688452 (this is discussed further in Sec. 2.9.2)
1	34–43	*First time derivative of the mean motion* (rev/day^2): 0.00000307
2	9–16	*Inclination* (degrees): 98.6328
2	18–25	*Right ascension of the ascending node* (degrees): 251.5324
2	27–33	*Eccentricity* (leading decimal point assumed): 0011501
2	35–42	*Argument of perigee* (degrees): 113.5534
2	44–51	*Mean anomaly* (degrees): 246.6853
2	53–63	*Mean motion* (rev/day): 14.23304826
2	64–68	*Revolution number at epoch* (rev): 11,663

Kepler's third law gives

$$a = \left[\frac{\mu}{n_0^2}\right]^{1/3}$$

$$= \underline{\underline{7192.335 \text{ km}}}$$

2.7 Apogee and Perigee Heights

Although not specified as orbital elements, the apogee height and perigee height are often required. As shown in App. B, the length of the radius vectors at apogee and perigee can be obtained from the geometry of the ellipse:

$$r_a = a(1+e) \tag{2.5}$$

$$r_p = a(1-e) \tag{2.6}$$

In order to find the apogee and perigee heights, the radius of the earth must be subtracted from the radii lengths, as shown in the following example.

Example 2.3 Calculate the apogee and perigee heights for the orbital parameters given in Table 2.1. Assume a mean earth radius of 6371 km.

Solution From Table 2.1: $e = .0011501$ and from Example 2.1 $a = 7192.335$ km. Using Eqs. (2.5) and (2.6):

$$r_a = 7192.335(1 + 0.0011501)$$

$$= 7200.607 \text{ km}$$

$$r_p = 7192.335(1 - 0.0011501)$$

$$= 7184.063 \text{ km}$$

The corresponding heights are:

$$h_a = r_a - R$$
$$= \underline{\underline{829.6 \text{ km}}}$$

$$h_p = r_p - R$$
$$= \underline{\underline{813.1 \text{ km}}}$$

2.8 Orbit Perturbations

The *keplerian orbit* described so far is ideal in the sense that it assumes that the earth is a uniform spherical mass and that the only force acting is the centrifugal force resulting from satellite motion balancing the gravitational pull of the earth. In practice, other forces which can be significant are the gravitational forces of the sun and the moon and atmospheric drag. The gravitational pulls of sun and moon have negligible effect on low-orbiting satellites, but they do affect satellites in the geostationary orbit as described in Sec. 3.5. Atmospheric drag, on the other hand, has negligible effect on geostationary satellites but does affect low-orbiting earth satellites below about 1000 km.

2.8.1 Effects of a nonspherical earth

For a spherical earth of uniform mass, Kepler's third law (Eq. 2.2) gives the nominal mean motion n_0 as

$$n_0 = \sqrt{\frac{\mu}{a^3}} \tag{2.7}$$

The 0 subscript is included as a reminder that this result applies for a perfectly spherical earth of uniform mass. However, it is known that the earth is not perfectly spherical, there being an equatorial bulge and a flattening at the poles, a shape described as an *oblate spheroid*. When the earth's oblateness is taken into account, the mean motion, denoted in this case by symbol n, is modified to (Wertz, 1984).

$$n = n_0\left[1 + \frac{K_1(1 - 1.5\sin^2 i)}{a^2(1 - e^2)^{1.5}}\right] \tag{2.8}$$

K_1 is a constant which evaluates to 66,063.1704 km². The earth's oblateness has negligible effect on the semimajor axis a, and if a is known, the mean motion is readily calculated. The orbital period taking into account the earth's oblateness is termed the *anomalistic period* (e.g., from perigee to perigee). The mean motion specified in the NASA bulletins is the reciprocal of the anomalistic period. The anomalistic period is

$$P_A = \frac{2\pi}{n} \text{ s} \tag{2.9}$$

where n is in radians per second.

If the known quantity is n (e.g., as is given in the NASA bulletins), one can solve Eq. (2.8) for a, keeping in mind that n_0 is also a function of a. Equation (2.8) may be solved for a by finding the root of the following equation:

$$n - \sqrt{\frac{\mu}{a^3}}\left[1 + \frac{K_1(1 - 1.5\sin^2 i)}{a^2(1 - e^2)^{1.5}}\right] = 0 \tag{2.10}$$

This is illustrated in the following example.

Example 2.4 A satellite is orbiting in the equatorial plane with a period from perigee to perigee of 12 h. Given that the eccentricity is 0.002, calculate the semimajor axis. The earth's equatorial radius is 6378.1414 km.

Solution Given data: $e = 0.002$; $i = 0°$; $P = 12$ h; $K_1 = 66063.1704$ km^2; $a_E =$ 6378.1414 km; $\mu = 3.986005 \times 1014$ m^3/s^2

The mean motion is:

$$n = \frac{2\pi}{P}$$
$$= 1.454 \times 10^{-4} \text{ s}^{-1}$$

Assuming this is the same as n_0, Kepler's third law gives

$$a = \left(\frac{\mu}{n^2}\right)^{1/3}$$
$$= \underline{\underline{26610 \text{ km}}}$$

Solving the root equation yields a value of 26,612 km.

The oblateness of the earth also produces two rotations of the orbital plane. The first of these, known as *regression of the nodes,* is where the nodes appear to slide along the equator. In effect, the line

of nodes, which is in the equatorial plane, rotates about the center of the earth. Thus Ω, the right ascension of the ascending node, shifts its position.

If the orbit is prograde (see Fig. 2.4), the nodes slide westward, and if retrograde, they slide eastward. As seen from the ascending node, a satellite in prograde orbit moves eastward, and in a retrograde orbit, westward. The nodes therefore move in a direction opposite to the direction of satellite motion, hence the term *regression of the nodes.* For a polar orbit ($i = 90°$), the regression is zero.

The second effect is rotation of apsides in the orbital plane, described below. Both effects depend on the mean motion n, the semimajor axis a, and the eccentricity e. These factors can be grouped into one factor K given by

$$K = \frac{nK_1}{a^2(1-e^2)^2} \tag{2.11}$$

K will have the same units as n. Thus, with n in rad/day, K will be in rad/day, and with n in degrees/day, K will be in degrees/day. An approximate expression for the rate of change of Ω with respect to time is (Wertz, 1984)

$$\frac{d\Omega}{dt} = -K\cos i \tag{2.12}$$

where i is the inclination. The rate of regression of the nodes will have the same units as n.

When the rate of change given by Eq. (2.12) is negative, the regression is westward, and when the rate is positive, the regression is eastward. It will be seen, therefore that for eastward regression, i must be greater than 90°, or the orbit must be retrograde. It is possible to choose values of a, e, and i such that the rate of rotation is 0.9856°/day eastward. Such an orbit is said to be sun synchronous and is described further in Sec. 2.10.

The other major effect produced by the equatorial bulge is a rotation of the line of apsides. This line rotates in the orbital plane, resulting in the argument of perigee changing with time. The rate of change is given by (Wertz, 1984)

$$\frac{d\omega}{dt} = K(2 - 2.5\sin^2 i) \tag{2.13}$$

Again, the units for the rate of rotation of the line of apsides will be the same as those for n (incorporated in K). When the inclination i is equal to 63.435°, the term within the parentheses is equal to zero, and

hence no rotation takes place. Use is made of this fact in the orbit chosen for the Russian Molniya satellites (see Probs. 2.23 and 2.24).

Denoting the epoch time by t_0, the right ascension of the ascending node by Ω_0, and the argument of perigee by ω_0 at epoch gives the new values for Ω and ω at time t as

$$\Omega = \Omega_0 + \frac{d\Omega}{dt}(t - t_0) \tag{2.14}$$

$$\omega = \omega_0 + \frac{d\omega}{dt}(t - t_0) \tag{2.15}$$

Keep in mind that the orbit is not a physical entity, and it is the forces resulting from an oblate earth, which act on the satellite to produce the changes in the orbital parameters. Thus, rather than follow a closed elliptical path in a fixed plane, the satellite drifts as a result of the regression of the nodes, and the latitude of the point of closest approach (the perigee) changes as a result of the rotation of the line of apsides. With this in mind, it is permissible to visualize the satellite as following a closed elliptical orbit but with the orbit itself moving relative to the earth as a result of the changes in Ω and ω. Thus, as stated earlier, the period P_A is the time required to go around the orbital path from perigee to perigee, even though the perigee has moved relative to the earth.

Suppose, for example, that the inclination is 90° so that the regression of the nodes is zero (from Eq. 2.12), and the rate of rotation of the line of apsides is $-K/2$ (from Eq. 2.13), and further, imagine the situation where the perigee at the start of observations is exactly over the ascending node. One period later the perigee would be at an angle $-KP_A/2$ relative to the ascending node or, in other words, would be south of the equator. The time between crossings at the *ascending node* would be $P_A(1 + K/2n)$, which would be the period observed from the earth. Recall that K will have the same units as n, for example, rad/s.

Example 2.5 Determine the rate of regression of the nodes and the rate of rotation of the line of apsides for the satellite parameters specified in Table 2.1. The value for a obtained in Example 2.2 may be used.

Solution From Table 2.1 and Example 2.2 $i = 98.6328°$; $e = 0.0011501$; NN = 14.23304826 day^{-1}; $a = 7192.335$ km, and the known constant: $K_1 = 66063.1704$ km^2

Converting n to rad/s:

$$n = 2\pi \text{NN}$$
$$= 0.00104 \text{ rad/s}$$

From Eq. (2.11):

$$K = \frac{nK_1}{a^2(1 - e^2)^2}$$
$$= 6.544 \text{ deg/day}$$

From Eq. (2.12):

$$\frac{d\Omega}{dt} = -K\cos i$$
$$= 0.981 \text{ deg/day}$$

From Eq. (2.13):

$$\frac{d\omega}{dt} = K(2 - 2.5\sin^2 i)$$
$$= \underline{\underline{-2.904 \text{ deg/day}}}$$

Example 2.6 Calculate, for the satellite in Example 2.5, the new values for ω and Ω one period after epoch.

Solution From Table 2.1:

$$\text{NN} = 14.23304826 \text{ day}^{-1};\ \omega_0 = 113.5534°;\ \Omega_0 = 251.5324°$$

The anomalistic period is

$$P_A = \frac{1}{\text{NN}}$$
$$= 0.070259 \text{ day}$$

This is also the time difference $(t - t_0)$ since the satellite has completed one revolution from perigee to perigee. Hence:

$$\Omega = \Omega_0 + \frac{d\Omega}{dt}(t - t_0)$$
$$= 251.5324 + 0.981(0.070259)$$
$$= \underline{\underline{251.601°}}$$

$$\omega = \omega_0 + \frac{d\omega}{dt}(t - t_0)$$
$$= 113.5534 + (-2.903)(0.070259)$$
$$= \underline{\underline{113.349°}}$$

In addition to the equatorial bulge, the earth is not perfectly circular in the equatorial plane; it has a small eccentricity of the order of 10^{-5}. This is referred to as the *equatorial ellipticity*. The effect of the equatorial ellipticity is to set up a gravity gradient, which has a pronounced effect on satellites in geostationary orbit (Sec. 7.4). Very briefly, a satellite in geostationary orbit ideally should remain fixed relative to the earth. The gravity gradient resulting from the equatorial ellipticity causes the satellites in geostationary orbit to drift to one of two stable points, which coincide with the minor axis of the equatorial ellipse. These two points are separated by 180° on the equator and are at approximately 75° E longitude and 105° W longitude. Satellites in service are prevented from drifting to these points through station-keeping maneuvers, described in Sec. 7.4. Because old, out-of-service satellites eventually do drift to these points, they are referred to as "satellite graveyards." It may be noted that the effect of equatorial ellipticity is negligible on most other satellite orbits.

2.8.2 Atmospheric drag

For near-earth satellites, below about 1000 km, the effects of atmospheric drag are significant. Because the drag is greatest at the perigee, the drag acts to reduce the velocity at this point, with the result that the satellite does not reach the same apogee height on successive revolutions.

The result is that the semimajor axis and the eccentricity are both reduced. Drag does not noticeably change the other orbital parameters, including perigee height. In the program used for generating the orbital elements given in the NASA bulletins, a pseudo-drag term is generated, which is equal to one-half the rate of change of mean motion (ADC USAF, 1980). An approximate expression for the change of major axis is

$$a \cong a_0 \left[\frac{n_0}{n_0 + n'_0(t - t_0)} \right]^{2/3} \tag{2.16}$$

where the "0" subscripts denote values at the reference time t_0, and n'_0 is the first derivative of the mean motion. The mean anomaly is also changed, an approximate value for the change being:

$$\delta M = \frac{n'_0}{2}(t - t_0)^2 \tag{2.17}$$

From Table 2.1 it is seen that the first time derivative of the mean motion is listed in columns 34–43 of line 1 of the NASA bulletin. For the

example shown in Fig. 2.6, this is 0.00000307 rev/day^2. Thus the changes resulting from the drag term will be significant only for long time intervals, and for present purposes will be ignored. For a more accurate analysis, suitable for long-term predictions, the reader is referred to ADC USAF (1980).

2.9 Inclined Orbits

A study of the general situation of a satellite in an inclined elliptical orbit is complicated by the fact that different parameters relate to different reference frames. The orbital elements are known with reference to the plane of the orbit, the position of which is fixed (or slowly varying) in space, while the location of the earth station is usually given in terms of the local geographic coordinates which rotate with the earth. Rectangular coordinate systems are generally used in calculations of satellite position and velocity in space, while the earth station quantities of interest may be the azimuth and elevation angles and range. Transformations between coordinate systems are therefore required.

Here, in order to illustrate the method of calculation for elliptical inclined orbits, the problem of finding the earth station look angles and range will be considered. It should be kept in mind that with inclined orbits the satellites are not geostationary, and therefore, the required look angles and range will change with time. Detailed and very readable treatments of orbital properties in general will be found, for example, in Bate et al. (1971) and Wertz (1984). Much of the explanation and the notation in this section is based on these two references.

Determination of the look angles and range involves the following quantities and concepts:

1. The *orbital elements*, as published in the NASA bulletins and described in Sec. 2.6
2. Various measures of *time*
3. The perifocal *coordinate system*, which is based on the orbital plane
4. The *geocentric-equatorial coordinate system*, which is based on the earth's equatorial plane
5. The *topocentric-horizon coordinate system*, which is based on the observer's horizon plane.

The two major coordinate transformations needed are:

- The satellite position measured in the perifocal system is transformed to the geocentric-horizon system in which the earth's rotation

is measured, thus enabling the satellite position and the earth station location to be coordinated.

- The satellite-to-earth station position vector is transformed to the topocentric-horizon system, which enables the look angles and range to be calculated.

2.9.1 Calendars

A calendar is a time-keeping device in which the year is divided into months, weeks, and days. Calendar days are units of time based on the earth's motion relative to the sun. Of course, it is more convenient to think of the sun moving relative to the earth. This motion is not uniform, and so a fictitious sun, termed the *mean sun*, is introduced.

The mean sun does move at a uniform speed but otherwise requires the same time as the real sun to complete one orbit of the earth, this time being the *tropical year*. A day measured relative to this mean sun is termed a *mean solar day*. Calendar days are mean solar days, and generally they are just referred to as days.

A tropical year contains 365.2422 days. In order to make the calendar year, also referred to as the *civil year*, more easily usable, it is normally divided into 365 days. The extra 0.2422 of a day is significant, and for example, after 100 years, there would be a discrepancy of 24 days between the calendar year and the tropical year. Julius Caesar made the first attempt to correct the discrepancy by introducing the *leap year*, in which an extra day is added to February whenever the year number is divisible by 4. This gave the *Julian calendar*, in which the civil year was 365.25 days on average, a reasonable approximation to the tropical year.

By the year 1582, an appreciable discrepancy once again existed between the civil and tropical years. Pope Gregory XIII took matters in hand by abolishing the days October 5 through October 14, 1582, to bring the civil and tropical years into line and by placing an additional constraint on the leap year in that years ending in two zeros must be divisible by 400 without remainder to be reckoned as leap years. This dodge was used to miss out 3 days every 400 years. To see this, let the year be written as $X00$ where X stands for the hundreds. For example, for 1900, $X = 19$. For $X00$ to be divisible by 400, X must be divisible by 4. Now a succession of 400 years can be written as $X + (n - 1)$, $X + n$, $X + (n + 1)$, and $X + (n + 2)$, where n is any integer from 0 to 9. If $X + n$ is evenly divisible by 4, then the adjoining three numbers are not, since some fraction from $-1/4$ to $2/4$ remains, so these three years would have to be omitted. The resulting calendar is the *Gregorian calendar*, which is the one in use today.

Example 2.7 Calculate the average length of the civil year in the Gregorian calendar.

Solution The nominal number of days in a 400-year period is 400 × 365 = 146,000. The nominal number of leap years is 400/4 = 100, but as shown earlier, this must be reduced by 3, and therefore, the number of days in 400 years of the Gregorian calendar is 146,000 + 100 − 3 = 146,097. This gives a yearly average of 146,097/400 = 365.2425.

In calculations requiring satellite predictions, it is necessary to determine whether a year is a leap year or not, and the simple rule is: If the year number ends in two zeros and is divisible by 400 without remainder, it is a leap year. Otherwise, if the year number is divisible by 4 without remainder, it is a leap year.

Example 2.8 Determine which of the following years are leap years: (*a*) 1987, (*b*) 1988, (*c*) 2000, (*d*) 2100.

Solution

(*a*) 1987/4 = 496.75 (therefore, 1987 is not a leap year)

(*b*) 1988/4 = 497 (therefore, 1988 is a leap year)

(*c*) 2000/400 = 5 (therefore, 2000 is a leap year)

(*d*) 2100/400 = 5.25 (therefore, 2100 is not a leap year, even though 2100 is divisible by 4 without remainder)

2.9.2 Universal time

Universal time coordinated (UTC) is the time used for all civil time–keeping purposes, and it is the time reference which is broadcast by the National Bureau of Standards as a standard for setting clocks. It is based on an atomic time-frequency standard. The fundamental unit for UTC is the mean solar day (see App. J in Wertz, 1984). In terms of "clock time," the mean solar day is divided into 24 h, an hour into 60 min, and a minute into 60 s. Thus there are 86,400 "clock seconds" in a mean solar day. Satellite-orbit epoch time is given in terms of UTC.

Example 2.9 Calculate the time in days, hours, minutes, and seconds for the epoch day 324.95616765.

Solution This represents the 324th day of the year plus 0.95616765 mean solar day. The decimal fraction in hours is 24 × 0.95616765 = 22.9480236; the decimal fraction of this expressed in minutes is 0.9480236 × 60 = 56.881416; the decimal fraction of this expressed in seconds is 0.881416 × 60 = 52.88496. Thus, the epoch is day 324, at 22 h, 58 m, 52.88 s.

Universal time coordinated is equivalent to *Greenwich mean time* (GMT), as well as *Zulu* (Z) time. There are a number of other "universal time" systems, all interrelated (Wertz, 1984) and all with the mean solar day as the fundamental unit. For present purposes, the distinction between these systems is not critical, and the term *universal time* (UT), will be used from now on.

For computations, UT will be required in two forms: as a fraction of a day and in degrees. Given UT in the normal form of hours, minutes, and seconds, it is converted to fractional days as

$$\mathrm{UT_{day}} = \frac{1}{24}\left(\text{hours} + \frac{\text{minutes}}{60} + \frac{\text{seconds}}{3600}\right) \tag{2.18}$$

In turn, this may be converted to degrees as

$$\mathrm{UT}^{\circ} = 360^{\circ} \times \mathrm{UT_{day}} \tag{2.19}$$

2.9.3 Julian dates*

Calendar times are expressed in UT, and although the time interval between any two events may be measured as the difference in their calendar times, the calendar time notation is not suited to computations where the timing of many events has to be computed. What is required is a reference time to which all events can be related in decimal days. Such a reference time is provided by the Julian zero time reference, which is 12 noon (12:00 UT) on January 1 in the year 4713 B.C.! Of course, this date would not have existed as such at the time; it is a hypothetical starting point, which can be established by counting backward according to a certain formula. For details of this intriguing time reference, see Wertz (1984, p. 20). The important point is that ordinary calendar times are easily converted to Julian dates, measured on a continuous time scale of Julian days. To do this, first determine the day of the year, keeping in mind that day zero, denoted as Jan 0.0 is midnight between December 30 and 31 of the previous year. For example, noon on December 31 would be January 0.5, and noon on January 1 would be January 1.5. It may seem strange that the last day of December should be denoted as "day zero in January," but it will be seen that this makes the day count correspond to the actual calendar day.

A Fortran program for calculating the Julian day for any date and time is given in Wertz (1984, p. 20), and a general method is given in Duffett-Smith (1986, p. 9). Once the Julian day is known for a given reference date and time, the Julian day for any other time can be easily calculated by adding or subtracting the required day difference. Some "reference times" are listed in Table 2.2.

*It should be noted that the Julian date is not associated with the Julian calendar introduced by Julius Caesar.

TABLE 2.2 Some Reference Julian Dates

January 0.0	Julian day
1999	2451178.5
2000	2451543.5
2001	2451909.5
2002	2452274.5
2003	2452639.5
2004	2453004.5
2005	2453370.5
2006	2453735.5
2007	2454100.5
2008	2454465.5
2009	2454831.5
2010	2455196.5

For convenience in calculations the day number of the year is given in Table 2.3.

Example 2.10 Find the Julian day for 13 h UT on 18 December 2000.

Solution The year 2000 is a leap year, and from Table 2.3, December 18 is day number 335 + 18 = 353. This is for midnight December 17/18. UT = 13 h as a fraction of a day is 13/24 = 0.5416667. From Table 2.2, the Julian date for January 0.0, 2000 is 2451543.5, and therefore the required Julian date is 2451543.5 + 353 + 0.5416667 = 2451897.0417.

In Sec. 2.9.7, certain calculations require a time interval measured in *Julian centuries*, where a Julian century consists of 36,525 mean solar days. The time interval is reckoned from a reference time of January 0.5, 1900, which corresponds to 2,415,020 Julian days.

TABLE 2.3 Day Number for the Last Day of the Month

Date	Day number for start of day (midnight) Numbers in parentheses are for leap years
January 31	31
February 28 (29)	59 (60)
March 31	90 (91)
April 30	120.5 (121.5)
May 31	151 (152)
June 30	181 (182)
July 31	212 (213)
August 31	243 (244)
September 30	273 (274)
October 31	304 (305)
November 30	334 (335)
December 31	365 (366)

Denoting the reference time as JD_{ref}, the Julian century by JC, and the time in question by JD, then the interval in Julian centuries from the reference time to the time in question is given by

$$T = \frac{JD - JD_{ref}}{JC} \tag{2.20}$$

This is illustrated in the following example.

Example 2.11 Find the time in Julian centuries from the reference time January 0.5, 1900 to 13 h UT on 18 December 2000.

Solution JD_{ref} = 2415020 days; JC = 36525 days. From Example 2.10: JD = 2451897.0417 days. Equation (2.20) gives

$$T = \frac{2451897.0417 - 2415020}{36525}$$
$$= \underline{\underline{1.00963838}}$$

Note that the time units are days and T is dimensionless.

2.9.4 Sidereal time

Sidereal time is time measured relative to the fixed stars (Fig. 2.7). It will be seen that one complete rotation of the earth relative to the fixed stars is not a complete rotation relative to the sun. This is because the earth moves in its orbit around the sun.

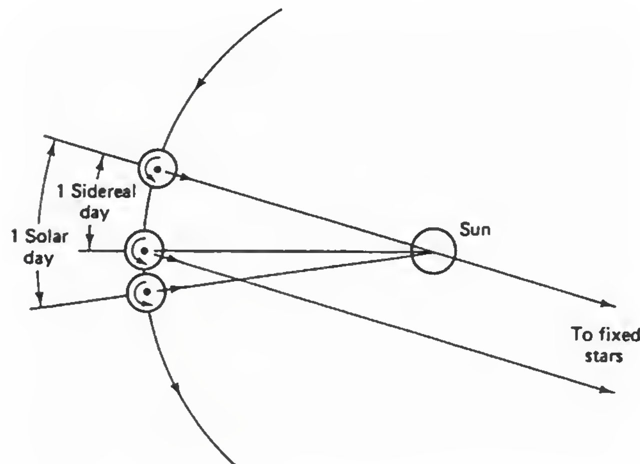


Figure 2.7 A sidereal day, or one rotation of the earth relative to fixed stars, is shorter than a solar day.

The *sidereal day* is defined as one complete rotation of the earth relative to the fixed stars. One sidereal day has 24 sidereal hours, 1 sidereal hour has 60 sidereal minutes, and 1 sidereal minute has 60 sidereal seconds. Care must be taken to distinguish between sidereal times and mean solar times, which use the same basic subdivisions. The relationships between the two systems, given in Bate et al. (1971), are

$$\begin{aligned} 1 \text{ mean solar day} &= 1.0027379093 \text{ mean sidereal days} \\ &= 24 \text{ h } 3 \text{ m } 56.55536 \text{ s sidereal time} \qquad (2.21) \\ &= 86{,}636.55536 \text{ mean sidereal seconds} \end{aligned}$$

$$\begin{aligned} 1 \text{ mean sidereal day} &= 0.9972695664 \text{ mean solar days} \\ &= 23 \text{ h } 56 \text{ m } 04.09054 \text{ s mean solar time} \qquad (2.22) \\ &= 86{,}164.09054 \text{ mean solar seconds} \end{aligned}$$

Measurements of longitude on the earth's surface require the use of sidereal time (discussed further in Sec. 2.9.7). The use of 23 h, 56 min as an approximation for the mean sidereal day will be used later in determining the height of the geostationary orbit.

2.9.5 The orbital plane

In the orbital plane, the position vector **r** and the velocity vector **v** specify the motion of the satellite, as shown in Fig. 2.8. For present purposes,

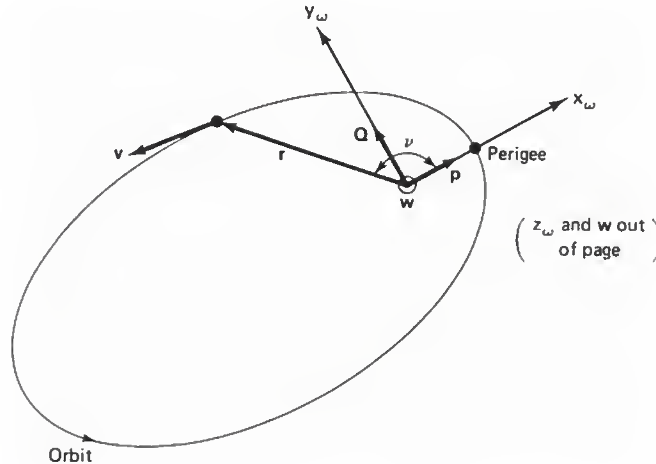


Figure 2.8 Perifocal coordinate system (**PQW** frame).

only the magnitude of the position vector is required. From the geometry of the ellipse (see App. B), this is found to be

$$r = \frac{a(1 - e^2)}{1 + e\cos\nu} \tag{2.23}$$

The true anomaly ν is a function of time, and determining it is one of the more difficult steps in the calculations.

The usual approach to determining ν proceeds in two stages. First, the mean anomaly M at time t is found. This is a simple calculation:

$$M = n(t - T_p) \tag{2.24}$$

Here, n is the mean motion, as previously defined in Eq. (2.8), and T_p is the time of perigee passage. The time of perigee passage T_p can be eliminated from Eq. (2.24) if one is working from the elements specified by NASA. For the NASA elements,

$$M_0 = n(t_0 - T_p)$$

Therefore,

$$T_p = t_0 - \frac{M_0}{n} \tag{2.25}$$

Substituting this in Eq. (2.24) gives

$$M = M_0 + n(t - t_0) \tag{2.26}$$

Consistent units must be used throughout. For example, with n in degrees/day, time $(t - t_0)$ must be in days and M_0 in degrees, and M will then be in degrees.

Example 2.12 Calculate the time of perigee passage for the NASA elements given in Table 2.1.

Solution The specified values at epoch are mean motion $n = 14.23304826$ rev/day, mean anomaly $M_0 = 246.6853°$, and $t_0 = 223.79688452$ days. In this instance it is only necessary to convert the mean motion to degrees/day, which is $360n$. Applying Eq. (2.25) gives

$$T = 223.7968452 - \frac{246.6853}{14.23304826 \times 360}$$
$$= \underline{\underline{223.74874044 \text{ days}}}$$

Once the mean anomaly M is known, the next step is to solve an equation known as *Kepler's equation*. Kepler's equation is formulated

in terms of an intermediate variable E, known as the *eccentric anomaly,* and is usually stated as

$$M = E - e\sin E \tag{2.27}$$

Kepler's equation is derived in App. B. This rather innocent looking equation is solved by iterative methods, usually by finding the root of the equation:

$$M - (E - e\sin E) = 0 \tag{2.28}$$

The following example shows how to solve for E graphically.

Example 2.13 Given that the mean anomaly is 205° and the eccentricity 0.0025, calculate the eccentric anomaly.

Solution The magnitude of the term $e \sin E$ will be much less than one, and therefore from Eq. (2.27) E will be approximately equal to M. Since M is greater than 180° the sin of E will be negative, and again from Eq. (2.27) this means that E will be smaller than M. Denote the second term on the left-hand side by $f(E) = E - e \sin E$; this can be evaluated for a range of values of E as shown in the following table. Writing Eq. (2.28) as $M - f(E) = 0$, the left-hand side of this can be evaluated also as shown in the table.

E (deg)	204.9	204.92	204.94	204.96	204.98	205
$f(E)$ rad	3.5772	3.5776	3.5779	3.5783	3.5786	3.579
$M - f(E)$ deg	0.04	0.02	−0.0004	−0.02	−0.04	−0.61

From this it is seen that $M - f(E) = 0$ occurs at about $E = 204.94°$.

Once E is found, ν can be found from an equation known as *Gauss' equation*, which is

$$\tan\frac{\nu}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{E}{2} \tag{2.29}$$

Gauss' equation is derived in App. B. Another result derived in App. B, which is useful for calculating the magnitude of the radius vector **r** as a function of E is

$$r = a(1 - e\cos E) \tag{2.30}$$

For near-circular orbits where the eccentricity is small, an approximation for ν directly in terms of M is

$$v \cong M + 2e\sin M + \frac{5}{4}e^2\sin 2M \tag{2.31}$$

Note that the first M term on the right-hand side must be in radians. This will give ν *in radians.*

Example 2.14 For satellite no. 14452 the NASA prediction bulletin for a certain epoch gives the eccentricity as 9.5981×10^{-3} and the mean anomaly as 204.9779°. The mean motion is 14.2171404 rev/day. Calculate the true anomaly and the magnitude of the radius vector 5 s after epoch. The semimajor axis is known to be 7194.9 km.

Solution The rotation in radians per second is

$$n = \frac{14.2171404 \times 2\pi}{86400}$$
$$\cong 0.001034 \text{ rad/s}$$

The mean anomaly of 204.9779°, in radians is 3.57754, and 5 s after epoch the mean anomaly becomes

$$M = 3.57754 + 0.001034 \times 5$$
$$\cong 3.5827 \text{ rad}$$
$$\nu \cong 3.5827 + 2 \times 9.5981 \times 10^{-3} \times \sin 3.5827$$
$$+ \frac{5}{4} \times (9.5981 \times 10^{-3})^2 \times \sin(2 \times 3.5827)$$
$$= \underline{\underline{3.5746 \text{ rad}}} \quad \underline{\underline{(= 204.81°)}}$$

Applying Eq. (2.23) gives r as

$$r = \frac{7194.9 \times (1 - 9.5981^2) \times 10^{-6}}{1 + 9.5981 \times 10^{-3} \times \cos 204.81}$$
$$= \underline{\underline{7257.5 \text{ km}}}$$

The magnitude r of the position vector $\boldsymbol{r}$ may be calculated by either Eq. (2.23) or Eq. (2.30). It may be expressed in vector form in the *perifocal coordinate system*. Here, the orbital plane is the fundamental plane, and the origin is at the center of the earth (only earth-orbiting satellites are being considered). The positive x axis lies in the orbital plane and passes through the perigee. Unit vector **P** points along the positive x axis as shown in Fig. 2.8. The positive y axis is rotated 90° from the x axis in the orbital plane, in the direction of satellite motion, and the unit vector is shown as **Q.** The positive z axis is normal to the orbital plane such that coordinates xyz form a right-hand set, and the unit vector is shown as **W.**

The subscript ω is used to distinguish the xyz coordinates in this system, as shown in Fig. 2.8. The position vector in this coordinate system, which will be referred to as the **PQW** *frame*, is given by

$$\mathbf{r} = (r \cos \nu)\mathbf{P} + (r \sin \nu)\mathbf{Q} \qquad (2.32)$$

The perifocal system is convenient for describing the motion of the satellite in the orbital plane. If the earth were uniformly spherical, the perifocal coordinates would be fixed in space, that is, inertial. However, the equatorial bulge causes rotations of the perifocal coordinate system, as described in Sec. 2.8.1. These rotations are taken into account when the satellite position is transferred from perifocal coordinates to *geocentric-equatorial coordinates*, described in the next section.

Example 2.15 Using the values $r = 7257.5$ km and $\nu = 204.81°$ obtained in the previous example, express r in vector form in the perifocal coordinate system.

Solution

$$r_P = 7257.5 \times \cos 204.81$$
$$= -6587.7 \text{ km}$$
$$r_Q = 7257.5 \times \sin 204.81$$
$$= -3045.3 \text{ km}$$

Hence

$$\mathbf{r} = \underline{\underline{-6587.7\mathbf{P} - 3045.3\mathbf{Q} \text{ km}}}$$

2.9.6 The geocentric-equatorial coordinate system

The *geocentric-equatorial coordinate system* is an inertial system of axes, the reference line being fixed by the fixed stars. The reference line is the line of Aries described in Sec. 2.5. (The phenomenon known as the *precession of the equinoxes* is ignored here. This is a very slow rotation of this reference frame, amounting to approximately 1.396971° per Julian century, where a Julian century consists of 36,525 mean solar days.) The fundamental plane is the earth's equatorial plane. Figure 2.9 shows the part of the ellipse above the equatorial plane and the orbital angles Ω, ω, and i. It should be kept in mind that Ω and ω may be slowly varying with time, as shown by Eqs. (2.12) and (2.13).

The unit vectors in this system are labeled **I**, **J**, and **K**, and the coordinate system is referred to as the **IJK** *frame*, with positive **I** pointing along the line of Aries. The transformation of vector **r** from the **PQW** frame to the **IJK** frame is most easily expressed by matrix multiplication. If A is an $m \times n$ matrix and B is an $n \times p$ matrix, the product AB is an $m \times p$ matrix (details of matrix multiplication will be found in most good

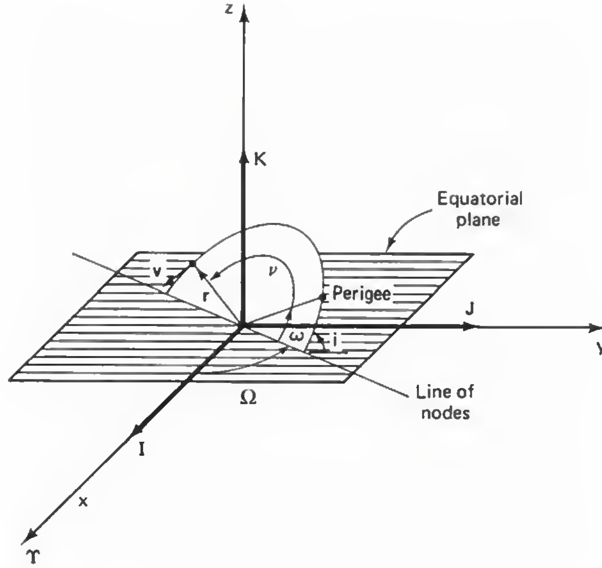


Figure 2.9 Geocentric-equatorial coordinate system (**IJK** frame).

texts on engineering mathematics. Many programmable calculators and computer programs have a built-in matrix multiplication function). The transformation matrix in this case is 3 × 2, given by Eq. (2.33*b*) and the **PQW** components form a 2 × 1 matrix. The components of *r* in the **IJK** frame appear as a 3 ×1 matrix given by

$$\begin{bmatrix} r_I \\ r_J \\ r_K \end{bmatrix} = \tilde{\mathbf{R}} \begin{bmatrix} r_P \\ r_Q \end{bmatrix} \tag{2.33a}$$

where the transformation matrix $\tilde{\mathbf{R}}$ is given by

$$\tilde{\mathbf{R}} = \begin{bmatrix} (\cos\Omega\cos\omega - \sin\Omega\sin\omega\cos i) & (-\cos\Omega\sin\omega - \sin\Omega\cos\omega\cos i) \\ (\sin\Omega\cos\omega + \cos\Omega\sin\omega\cos i) & (-\sin\Omega\sin\omega + \cos\Omega\cos\omega\cos i) \\ (\sin\omega\sin i) & (\cos\omega\sin i) \end{bmatrix} \tag{2.33b}$$

It should be noted that the angles Ω and ω take into account the rotations resulting from the earth's equatorial bulge, as described in Sec. 2.8.1. The matrix multiplication is most easily carried out by computer.

Example 2.16 Calculate the magnitude of the position vector in the **PQW** frame for the orbit specified below. Calculate also the position vector in the **IJK** frame and its magnitude. Confirm that this remains unchanged from the value obtained in the **PQW** frame.

Solution The given orbital elements are

$$\Omega = 300°;\ \omega = 60°;\ i = 65°;\ r_P = -6500 \text{ km};\ r_Q = 4000 \text{ km}$$

The magnitude of the **r** vector is

$$r = \sqrt{(-6500)^2 + (4000)^2}$$
$$= \underline{\underline{7632.2 \text{ km}}}$$

Substituting the angle values in Eq. (2.33b) gives

$$\tilde{\mathbf{R}} = \begin{bmatrix} 0.567 & -0.25 \\ -0.25 & 0.856 \\ 0.785 & 0.453 \end{bmatrix}$$

The vector components in the **IJK** frame are

$$\begin{bmatrix} r_I \\ r_J \\ r_K \end{bmatrix} = \begin{bmatrix} 0.567 & -0.25 \\ -0.25 & 0.856 \\ 0.785 & 0.453 \end{bmatrix} \begin{bmatrix} -6500 \\ 4000 \end{bmatrix}$$
$$= \begin{bmatrix} -4685.3 \\ 5047.7 \\ -3289.1 \end{bmatrix} \text{km}$$

The magnitude of the **r** vector in the **IJK** frame is

$$\mathbf{r} = \sqrt{(-4685.3)^2 + (5047.7)^2 + (-3289.1)^2}$$
$$= \underline{\underline{7632.2 \text{ km}}}$$

This is seen to be the same as that obtained from the **P** and **Q** components.

2.9.7 Earth station referred to the IJK frame

The earth station's position is given by the geographic coordinates of latitude λ_E and longitude ϕ_E. (Unfortunately, there does not seem to be any standardization of the symbols used for latitude and longitude. In some texts, as here, the Greek lambda is used for latitude and the Greek phi for longitude. In other texts, the reverse of this happens. One minor advantage of the former is that latitude and lambda both begin with the same "la" which makes the relationship easy to remember.)

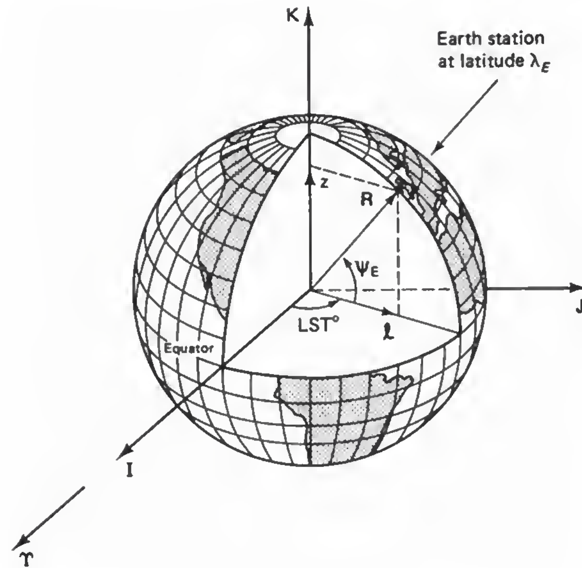


Figure 2.10 Position vector **R** of the earth relative to the **IJK** frame.

Care also must be taken regarding the sign conventions used for latitude and longitude because different systems are sometimes used, depending on the application. In this book, north latitudes will be taken as positive numbers and south latitudes as negative numbers, zero latitude, of course, being the equator. Longitudes east of the Greenwich meridian will be taken as positive numbers, and longitudes west, as negative numbers.

The position vector of the earth station relative to the **IJK** frame is **R** as shown in Fig. 2.10. The angle between **R** and the equatorial plane, denoted by ψ_E in Fig. 2.10, is closely related, but not quite equal to, the earth station latitude. More will be said about this angle shortly. **R** is obviously a function of the rotation of the earth, and so first it is necessary to find the position of the Greenwich meridian relative to the **I** axis as a function of time. The angular distance from the **I** axis to the Greenwich meridian is measured directly as *Greenwich sidereal time* (GST), also known as the *Greenwich hour angle*, or GHA. Sidereal time is described in Sec. 2.9.4.

GST may be found using values tabulated in some almanacs (see Bate et al., 1971), or it may be calculated using formulas given in Wertz (1984). In general, sidereal time may be measured in time units of sidereal days, hours, minutes, seconds, or it may be measured in angular units (degrees, minutes, seconds, or radians). Conversion is easily accomplished, since 2π radians or 360° correspond to 24 sidereal hours. The formula for GST in degrees is

$$\mathrm{GST} = 99.9610° + 36000.7689° \times T + 0.0004° \times T^2 + \mathrm{UT}° \quad (2.34)$$

Here, UT° is universal time expressed in degrees, as given by Eq. (2.19). T is the time in Julian centuries, given by Eq. (2.20).

Once GST is known, the *local sidereal time* (LST) is found by adding the east longitude of the station in degrees. East longitude for the earth station will be denoted as EL. Recall that previously longitude was expressed in positive degrees east and negative degrees west. For east longitudes, EL $= \phi_E$, while for west longitudes, EL $= 360° + \phi_E$. For example, for an earth station at east longitude 40°, EL = 40°. For an earth station at west longitude 40°, EL = 360 + (−40) = 320°. Thus the LST in degrees is given by

$$\text{LST} = \text{GST} + \text{EL} \tag{2.35}$$

The procedure is illustrated in the following examples

Example 2.17 Find the GST for 13 h UT on 18 December 2000.

Solution From Example 2.11: $T = 1.00963838$. The individual terms of Eq. (2.34) are:

$$X = 36000.7689° \times T = 347.7578° \quad (\text{mod } 360°)$$

$$Y = 0.0004° \times T^2 = 0.00041° \quad (\text{mod } 360°)$$

$$\text{UT} = \frac{13}{24} \times 360° = 195°$$

$$\text{GST} = 99.6910° + X + Y + \text{UT}$$

$$= \underline{\underline{282.4493° \ (\text{mod } 360°)}}$$

Example 2.18 Find the LST for Thunder Bay, longitude 89.26°W for 13 h UT on December 18, 2000.

Solution Expressing the longitude in degrees west: WL = −89.26°

In degrees east this is EL = 360° + (−89.26°) = 270.74°

$$\text{LST} = \text{GST} + \text{EL}$$

$$= 282.449 + 270.74$$

$$= \underline{\underline{93.189° \ (\text{mod } 360°)}}$$

Knowing the LST enables the position vector **R** of the earth station to be located with reference to the **IJK** frame as shown in Fig. 2.10. However, when **R** is resolved into its rectangular components, account must be taken of the oblateness of the earth. The earth may be modeled as an *oblate spheroid,* in which the equatorial plane is circular, and any

meridional plane (i.e., any plane containing the earth's polar axis) is elliptical, as illustrated in Fig. 2.11. For one particular model, known as a *reference ellipsoid,* the semimajor axis of the ellipse is equal to the equatorial radius, the semiminor axis is equal to the polar radius, and the surface of the ellipsoid represents the *mean sea level.* Denoting the semimajor axis by a_E and the semiminor axis by b_E and using the known values for the earth's radii gives

$$a_E = 6378.1414 \text{ km} \tag{2.36}$$

$$b_E = 6356.755 \text{ km} \tag{2.37}$$

From these values the eccentricity of the earth is seen to be

$$e_E = \frac{\sqrt{a_E^2 - b_E^2}}{a_E} = 0.08182 \tag{2.38}$$

In Figs. 2.10 and 2.11, what is known as the *geocentric latitude* is shown as ψ_E. This differs from what is normally referred to as latitude. An imaginary plumb line dropped from the earth station makes an angle λ_E with the equatorial plane, as shown in Fig. 2.11. This is known as the *geodetic latitude,* and for all practical purposes here, this can be taken as the geographic latitude of the earth station.

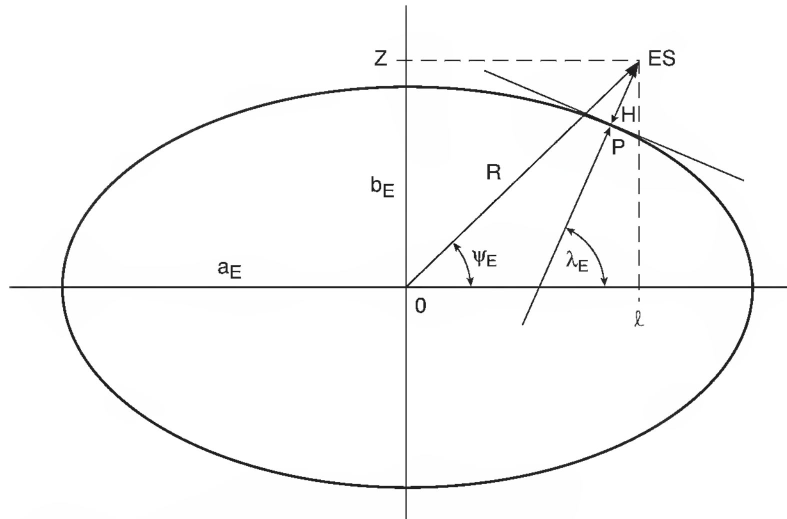


Figure 2.11 Reference ellipsoid for the earth showing the geocentric latitude ψ_E and the geodetic latitude λ_E.

With the height of the earth station above mean sea level denoted by H, the geocentric coordinates of the earth station position are given in terms of the geodetic coordinates by (Thompson, 1966)

$$N = \frac{a_E}{\sqrt{1 - e_E^2 \sin^2 \lambda_E}} \tag{2.39}$$

$$R_I = (N + H)\cos\lambda_E \cos \mathrm{LST}$$
$$= l \cos \mathrm{LST} \tag{2.40}$$

$$R_J = (N + H)\cos\lambda_E \sin \mathrm{LST}$$
$$= l \sin \mathrm{LST} \tag{2.41}$$

$$R_K = \left[N\left(1 - e_E^2\right) + H\right]\sin\lambda_E$$
$$= z \tag{2.42}$$

Example 2.19 Find the components of the radius vector to the earth station at Thunder Bay, given that the latitude is 48.42°, the height above sea level is 200 m, and the LST is 167.475°.

Solution With all distances in km, $H = 0.2$ km

$$N = \frac{a_E}{\sqrt{1 - e_E^2 \sin^2 \lambda_E}}$$

$$= \frac{6378.1414}{\sqrt{1 - 0.08182^2 \sin^2 48.42}}$$

$$= 6390.121 \text{ km}$$

$$l = (N + H)\cos\lambda_E$$

$$= 6390.321 \times 0.66366$$

$$= 4241.033 \text{ km}$$

$$R_I = l\cos(\mathrm{LST})$$

$$= 4241.033 \times (-0.9762)$$

$$= \underline{\underline{-4140.103 \text{ km}}}$$

$$R_J = l\sin(\mathrm{LST})$$

$$= 4241.033 \times (0.216865)$$

$$= \underline{\underline{919.734 \text{ km}}}$$

$$R_K = [N(1 - e_E^2) + H]\sin\lambda_E$$

$$= [6390.121 \times (1 - 0.08182^2) + 0.2] \times 0.74803$$

$$= \underline{\underline{4748.151 \text{ km}}}$$

At this point, both the satellite radius vector **r** and the earth station radius vector **R** are known in the **IJK** frame for any position of satellite and earth. From the vector diagram shown in Fig. 2.12*a*, the range vector ρ is obtained as

$$\boldsymbol{\rho} = \mathbf{r} - \mathbf{R} \tag{2.43}$$

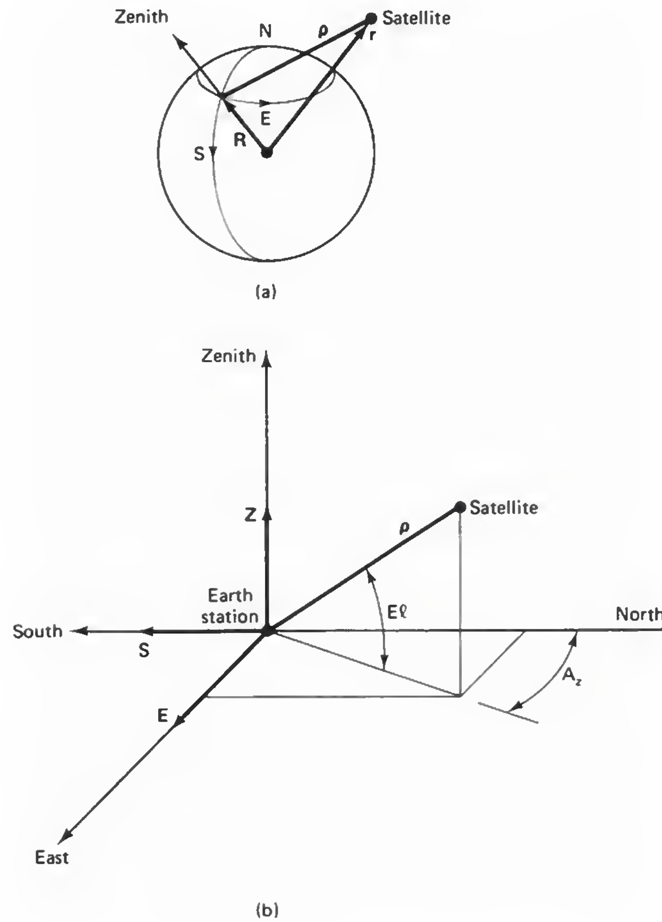


Figure 2.12 Topocentric-horizon coordinate system (**SEZ** frame): (*a*) overall view; (*b*) detailed view.

This gives $\boldsymbol{\rho}$ in the **IJK** frame. It then remains to transform $\boldsymbol{\rho}$ to the observer's frame, known as the *topocentric-horizon frame*, shown in Fig. 2.12*b*.

2.9.8 The topocentric-horizon coordinate system

The position of the satellite, as measured from the earth station, is usually given in terms of the azimuth and elevation angles and the range ρ. These are measured in the *topocentric-horizon coordinate system* illustrated in Fig. 2.12*b*. In this coordinate system, the fundamental plane is the observer's horizon plane. In the notation given in Bate et al. (1971), the positive x axis is taken as south, the unit vector being denoted by **S.** The positive y axis points east, the unit vector being **E.** The positive z axis is "up," pointing to the observer's zenith, the unit vector being **Z.** (*Note*: This is not the same z as that used in Sec. 2.9.7.) The frame is referred to as the **SEZ** frame, which of course rotates with the earth.

As shown in the previous section, the range vector $\boldsymbol{\rho}$ is known in the **IJK** frame, and it is now necessary to transform this to the **SEZ** frame. Again, this is a standard transformation procedure. See Bate et al. (1971).

$$\begin{bmatrix} \rho_S \\ \rho_E \\ \rho_Z \end{bmatrix} = \begin{bmatrix} \sin\psi_E \cos \mathrm{LST} & \sin\psi_E \sin \mathrm{LST} & -\cos\psi_E \\ -\sin \mathrm{LST} & \cos \mathrm{LST} & 0 \\ \cos\psi_E \cos \mathrm{LST} & \cos\psi_E \sin \mathrm{LST} & \sin\psi_E \end{bmatrix} \begin{bmatrix} \rho_I \\ \rho_J \\ \rho_K \end{bmatrix} \quad (2.44)$$

From Fig. 2.11, the geocentric angle ψ_E is seen to be given by

$$\psi_E = \arctan\frac{z}{l} \quad (2.45)$$

The coordinates l and z given in Eqs. (2.40) and (2.42) are known in terms of the earth station height and latitude, and hence the range vector is known in terms of these quantities and the LST. As a point of interest, for zero height, the angle ψ_E is related to λ_E by

$$\tan\psi_{E(H=0)} = \left(1 - e_E^2\right)\tan\lambda_E \quad (2.46)$$

Here, e_E is the earth's eccentricity, equal to 0.08182. The difference between the geodetic and geocentric latitudes reaches a maximum at a geocentric latitude of 45°, when the geodetic latitude is 45.192°.

The magnitude of the range is

$$\rho = \sqrt{\rho_S^2 + \rho_E^2 + \rho_Z^2} \quad (2.47)$$

TABLE 2.4 Azimuth Angles

ρ_S	ρ_E	Azimuth (degrees)
−	+	α
+	+	$180° - \alpha$
+	−	$180° + \alpha$
−	−	$360° - \alpha$

The antenna elevation angle is

$$El = \arcsin \frac{\rho_Z}{\rho} \tag{2.48}$$

The antenna azimuth angle is found from

$$\alpha = \arctan \frac{|\rho_E|}{|\rho_S|} \tag{2.49}$$

The azimuth depends on which quadrant α is in. With α in degrees the azimuth is as given in Table 2.4.

Example 2.20 The **IJK** range vector components for a certain satellite, at GST = 240°, are as given below. Calculate the corresponding range and the look angles for an earth station the coordinates for which are—latitude 48.42°N, longitude 89.26°W, height above mean sea level 200 m.

Solution Given data: $\rho_I = -1280$ km; $\rho_J = -1278$ km; $\rho_K = 66$ km; GST = 240°; $\lambda_E = 48.42°$; $\phi_E = -89.26°$; $H = 200$ m

The required earth constants are $a_E = 6378.1414$ km; $e_E = 0.08182$

$$\begin{aligned}
N &= \frac{a_E}{\sqrt{1 - e_E^2 \sin^2 \lambda_E}} \\
&= \frac{6378.1414}{\sqrt{1 - 0.08182^2 \sin^2 48.42}} \\
&= 6390.121 \text{ km} \\
l &= (N + H) \cos \lambda_E \\
&= 6390.321 \times 0.66366 \\
&= 4241.033 \text{ km} \\
z &= \left[N\left(1 - e_E^2\right) + H\right] \sin \lambda_E \\
&= [6390.121 \times (1 - 0.08182^2) + 0.2] \times 0.74803 \\
&= \underline{\underline{4748.151 \text{ km}}}
\end{aligned}$$

$$\psi_E = \arctan\frac{z}{l}$$
$$= 42.2289°$$

Substituting the known values in Eq. (2.44), and with all distances in km:

$$\begin{bmatrix}\rho_S\\ \rho_E\\ \rho_Z\end{bmatrix} = \begin{bmatrix}-.6507 & .3645 & -.6662\\ -.4888 & -.8724 & 0\\ -.5812 & .3256 & .7458\end{bmatrix}\begin{bmatrix}-1280\\ -1278\\ 66\end{bmatrix}$$

$$= \begin{bmatrix}322.9978\\ 1740.571\\ 376.9948\end{bmatrix}\text{km}$$

The magnitude is

$$\rho = \sqrt{322.9978^2 + 1740.571^2 + 376.9948^2}$$
$$\cong \underline{\underline{1810\text{ km}}}$$

The antenna angle of elevation is

$$El = \arcsin\frac{376.9948}{1810}$$
$$\cong \underline{\underline{12°}}$$

The angle α is

$$\alpha = \arctan\frac{1740.571}{322.9978}$$
$$= 79.487°$$

Since both ρ_E and ρ_S are positive, Table 2.4 gives the azimuth as

$$Az = 180° - \alpha$$
$$= \underline{\underline{100.5°}}$$

2.9.9 The subsatellite point

The point on the earth vertically under the satellite is referred to as the *subsatellite point*. The latitude and longitude of the subsatellite point and the height of the satellite above the subsatellite point can be determined from knowledge of the radius vector **r**. Figure 2.13 shows the meridian plane which cuts the subsatellite point. The height of the terrain above the reference ellipsoid at the subsatellite point is denoted by H_{SS}, and the height of the satellite above this, by h_{SS}. Thus the total height of the

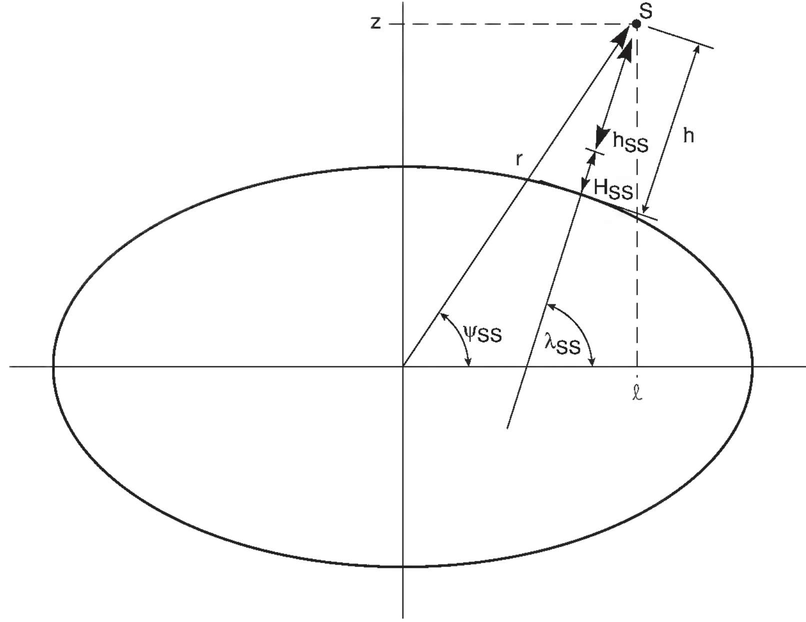


Figure 2.13 Geometry for determining the subsatellite point.

satellite above the reference ellipsoid is

$$h = H_{SS} + h_{SS} \tag{2.50}$$

Now the components of the radius vector **r** in the **IJK** frame are given by Eq. (2.33). Figure 2.13 is seen to be similar to Fig. 2.11, with the difference that r replaces R, the height to the point of interest is h rather than H, and the subsatellite latitude λ_{SS} is used. Thus Eqs. (2.39) through (2.42) may be written for this situation as

$$N = \frac{a_E}{\sqrt{1 - e_E^2 \sin^2 \lambda_{SS}}} \tag{2.51}$$

$$r_I = (N + h)\cos\lambda_{SS}\cos \text{LST} \tag{2.52}$$

$$r_J = (N + h)\cos\lambda_{SS}\sin \text{LST} \tag{2.53}$$

$$r_K = \left[N\left(1 - e_E^2\right) + h\right]\sin\lambda_{SS} \tag{2.54}$$

We now have three equations in three unknowns, LST, λ_E, and h, which can be solved. In addition, by analogy with the situation shown in Fig. 2.10, the east longitude is obtained from Eq. (2.35) as

$$\text{EL} = \text{LST} - \text{GST} \tag{2.55}$$

where GST is the Greenwich sidereal time.

Example 2.21 Determine the subsatellite height, latitude, and LST for the satellite in Example 2.16.

Solution From Example 2.16, the known components of the radius vector **r** in the **IJK** frame can be substituted in the left-hand side of Eqs.(2.52) through (2.54). The known values of a_E and e_E can be substituted in the right-hand side to give

$$-4685.3 = \left(\frac{6378.1414}{\sqrt{1 - 0.08182^2 \sin^2 \lambda_{SS}}} + h\right) \cos \lambda_{SS} \cos \mathrm{LST}$$

$$5047.7 = \left(\frac{6378.1414}{\sqrt{1 - 0.08182^2 \sin^2 \lambda_{SS}}} + h\right) \cos \lambda_{SS} \sin \mathrm{LST}$$

$$-3289.1 = \left(\frac{6378.1414 \times (1 - 0.08182^2)}{\sqrt{1 - 0.08182^2 \sin^2 \lambda_{SS}}} + h\right) \cos \lambda_{SS} \cos \mathrm{LST}$$

Each equation contains the unknowns LST, λ_{ss}, and h. Unfortunately, these unknowns cannot be separated out in the form of explicit equations. The following values were obtained by a computer solution.

$$\lambda_{ss} \cong \underline{\underline{-25.654°}}$$

$$h \cong \underline{\underline{1258.012 \text{ km}}}$$

$$\mathrm{LST} \cong \underline{\underline{132.868°}}$$

2.9.10 Predicting satellite position

The basic factors affecting satellite position are outlined in the previous sections. The NASA two-line elements are generated by prediction models contained in Spacetrack report no. 3 (ADC USAF, 1980), which also contains Fortran IV programs for the models. Readers desiring highly accurate prediction methods are referred to this report. Spacetrack report No. 4 (ADC USAF, 1983) gives details of the models used for atmospheric density analysis.

2.10 Local Mean Solar Time and Sun-Synchronous Orbits

The *celestial sphere* is an imaginary sphere of infinite radius, where the points on the surface of the sphere represent stars or other celestial objects. The points represent directions, and distance has no significance for the sphere. The orientation and center of the sphere can be selected to suit the conditions being studied, and in Fig. 2.14 the

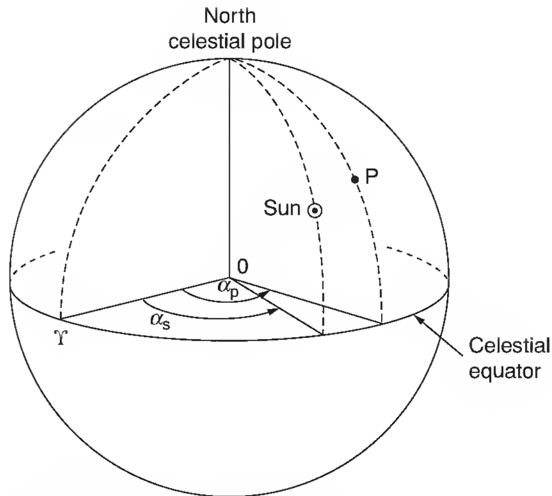


Figure 2.14 Sun-synchronous orbit.

sphere is centered on the geocentric-equatorial coordinate system (see Sec. 2.9.6). What this means is that the celestial equatorial plane coincides with the earth's equatorial plane, and the direction of the north celestial pole coincides with the earth's polar axis. For clarity the **IJK** frame is not shown, but from the definition of the line of Aries in Sec. 2.9.6, the point for Aries lies on the celestial equator where this is cut by the x-axis, and the z-axis passes through the north celestial pole.

Also shown in Fig. 2.14 is the sun's meridian. The angular distance along the celestial equator, measured eastward from the point of Aries to the sun's meridian is the *right ascension of the sun*, denoted by α_s. In general, the right ascension of a point P, is the angle, measured eastward along the celestial equator from the point of Aries to the meridian passing through P. This is shown as α_P in Fig. 2.14. The *hour angle* of a star is the angle measured westward along the celestial equator from the meridian to meridian of the star. Thus for point P the hour angle of the sun is $(\alpha_p - \alpha_s)$ measured westward (the hour angle is measured in the opposite direction to the right ascension).

Now the *apparent solar time* of point P is the local hour angle of the sun, expressed in hours, plus 12 h. The 12 h is added because zero hour angle corresponds to midday, when the P meridian coincides with the sun's meridian. Because the earth's path around the sun is elliptical rather than circular, and also because the plane containing the path of the earth's orbit around the sun (the *ecliptic plane*) is inclined at an angle of approximately 23.44°, the apparent solar time does not measure out

uniform intervals along the celestial equator, in other words, the length of a solar day depends on the position of the earth relative to the sun. To overcome this difficulty a fictitious mean sun is introduced, which travels in uniform circular motion around the sun (this is similar in many ways to the mean anomaly defined in Sec. 2.5). The time determined in this way is the *mean solar time*. Tables are available in various almanacs which give the relationship between mean solar time and apparent solar time through the *equation of time*.

The relevance of this to a satellite orbit is illustrated in Fig. 2.15. This shows the trace of a satellite orbit on the celestial sphere, (again keeping in mind that directions and not distances are shown). Point *A* corresponds to the ascending node. The hour angle of the sun from the ascending node of the satellite is $\Omega - \alpha_s$ measured westward. The hour angle of the sun from the satellite (projected to *S* on the celestial sphere) is $\Omega - \alpha_s + \beta$ and thus the local mean (solar) time is

$$t_{\mathrm{SAT}} = \frac{1}{15}(\Omega - \alpha_s + \beta) + 12 \tag{2.56}$$

To find β requires solving the spherical triangle defined by the points *ASB*. This is a right spherical triangle because the angle between the meridian plane through *S* and the equatorial plane is a right angle. The triangle also contains the inclination i (the angle between the orbital plane and the equatorial plane) and the latitude λ (the angle measured at the center of the sphere going north along the meridian through *S*). The inclination i and the latitude λ are the same angles already introduced in connection with orbits. The solution of the right spherical triangle

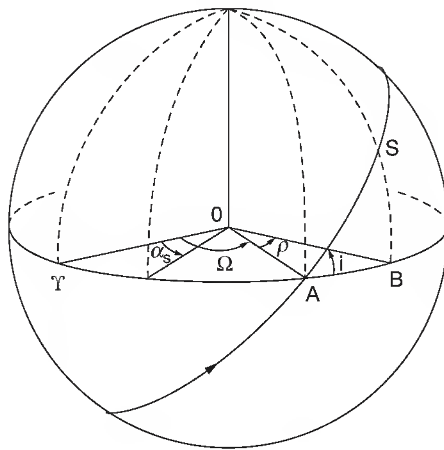


Figure 2.15 The condition for sun synchronicity is that the local solar time should be constant.

(see Wertz, 1984) yields for β

$$\beta = \arcsin\left(\frac{\tan\lambda}{\tan i}\right) \tag{2.57}$$

The local mean (solar) time for the satellite is therefore

$$t_{\text{SAT}} = \frac{1}{15}\left[\Omega - \alpha_s + \arcsin\left(\frac{\tan\lambda}{\tan i}\right)\right] + 12 \tag{2.58}$$

Notice that as the inclination i approaches 90° angle β approaches zero.

Accurate formulas are available for calculating the right ascension of the sun, but a good approximation to this is

$$\alpha_s = \frac{\Delta d}{365.24}360° \tag{2.59}$$

where Δd is the time in days from the vernal equinox. This is so because in one year of approximately 365.24 days the earth completes a 360° orbit around the sun.

For a *sun-synchronous orbit* the local mean time must remain constant. The advantage of a sun-synchronous orbit for weather satellites and environmental satellites is that the each time the satellite passes over a given latitude, the lighting conditions will be approximately the same. Eq. (2.58) shows that for a given latitude and fixed inclination, the only variables are α_s and Ω. In effect, the angle $(\Omega - \alpha_s)$ must be constant for a constant local mean time. Let Ω_0 represent the right ascension of the ascending node at the vernal equinox and Ω' the time rate of change of Ω then

$$\begin{aligned} t_{\text{SAT}} &= \frac{1}{15}\left[\Omega_0 + \Omega'\Delta d - \frac{\Delta d}{365.24}360° + \arcsin\left(\frac{\tan\lambda}{\tan i}\right)\right] + 12 \\ &= \frac{1}{15}\left[\Omega_0 + \left(\Omega' - \frac{360}{365.24}\right)\Delta d + \arcsin\left(\frac{\tan\lambda}{\tan i}\right)\right] + 12 \end{aligned} \tag{2.60}$$

For this to be constant the coefficient of Δd must be zero, or

$$\begin{aligned} \Omega' &= \frac{360°}{365.24} \\ &= 0.9856 \text{ degrees/day} \end{aligned} \tag{2.61}$$

Use is made of the regression of the nodes to achieve sun synchronicity. As shown in Sec. 2.8.1 by Eqs. (2.12) and (2.14), the rate of regression of the nodes and the direction are determined by the orbital elements

TABLE 2.5 Tiros-N Series Orbital Parameters

	833-km orbit	870-km orbit
Inclination	98.739°	98.899°
Nodal period	101.58 min	102.37 min
Nodal regression	25.40°/day E	25.59°/day E
Nodal precession	0.986°/day E	0.986°/day E
Orbits per day	14.18	14.07

SOURCE: Schwalb, 1982a and b.

a, e, and *i*. These can be selected to give the required regression of 0.9856° east per day. The orbital parameters for the Tiros-N satellites are listed in Table 2.5. These satellites follow near-circular, near-polar orbits.

2.11 Standard Time

Local mean time is not suitable for civil time–keeping purposes because it changes with longitude (and latitude), which would make it difficult to order day-to-day affairs. The approach taken internationally is to divide the world into 1-h time zones, the zonal meridians being 15° apart at the equator. The Greenwich meridian is used as zero reference and in the time zone that is ±7.5° about the Greenwich meridian the civil time is the same as the GMT. Care must be taken, however, since in the spring the clocks are advanced by 1 h, leading to *British summer time* (BST), also known as daylight saving time. Thus BST is equal to GMT plus 1 h.

In the first zone east of the GMT zone, the basic civil time is GMT + 1 h, and in the first zone west of the GMT zone, the basic civil time is GMT–1 h. One hour is added or subtracted for each additional zone east or west. Again, care must be taken to allow for summer time if it is in force (not all regions have the same summer time adjustment, and some regions may not use it at all). Also, in some instances the zonal meridians are adjusted where necessary to suit regional or country boundaries.

Orbital elements are normally specified in relation to GMT (or as noted in Sec. 2.9.2, UTC), but results (such as times of equatorial crossings) usually need to be known in the standard time for the zone where observations are being made. Care must be taken therefore to allow for the zone change, and for daylight saving time if in force. Many useful time zone maps and other information can be obtained from the Internet through a general search for "time zones."

2.12 Problems

2.1. State Kepler's three laws of planetary motion. Illustrate in each case their relevance to artificial satellites orbiting the earth.

2.2. Using the results of App. B, show that for any point P, the sum of the focal distances to S and S' is equal to $2a$.

2.3. Show that for the ellipse the differential element of area $dA = r^2 d\nu/2$, where $d\nu$ is the differential of the true anomaly. Using Kepler's second law, show that the ratio of the speeds at apoapsis and periapsis (or apogee and perigee for an earth-orbiting satellite) is equal to

$$(1 - e)/(1 + e)$$

2.4. A satellite orbit has an eccentricity of 0.2 and a semimajor axis of 10,000 km. Find the values of (*a*) the latus rectum; (*b*) the minor axis; (*c*) the distance between foci.

2.5. For the satellite in Prob. 2.4, find the length of the position vector when the true anomaly is 130°.

2.6. The orbit for an earth-orbiting satellite has an eccentricity of 0.15 and a semimajor axis of 9000 km. Determine (*a*) its periodic time; (*b*) the apogee height; (*c*) the perigee height. Assume a mean value of 6371 km for the earth's radius.

2.7. For the satellite in Prob. 2.6, at a given observation time during a south to north transit, the height above ground is measured as 2000 km. Find the corresponding true anomaly.

2.8. The semimajor axis for the orbit of an earth-orbiting satellite is found to be 9500 km. Determine the mean anomaly 10 min after passage of perigee.

2.9. The exact conversion factor between feet and meters is 1 ft = 0.3048 m. A satellite travels in an unperturbed circular orbit of semimajor axis a = 27,000 km. Determine its tangential speed in (*a*) km/s, (*b*) ft/s, and (*c*) mi/h.

2.10. Explain what is meant by *apogee height* and *perigee height*. The Cosmos 1675 satellite has an apogee height of 39,342 km and a perigee height of 613 km. Determine the semimajor axis and the eccentricity of its orbit. Assume a mean earth radius of 6371 km.

2.11. The Aussat 1 satellite in geostationary orbit has an apogee height of 35,795 km and a perigee height of 35,779 km. Assuming a value of 6378 km for the earth's equatorial radius, determine the semimajor axis and the eccentricity of the satellite's orbit.

2.12. Explain what is meant by the ascending and descending nodes. In what units would these be measured, and in general, would you expect them to change with time?

2.13. Explain what is meant by (*a*) line of apsides and (*b*) line of nodes. Is it possible for these two lines to be coincident?

2.14. With the aid of a neat sketch, explain what is meant by each of the angles: *inclination; argument of perigee; right ascension of the ascending node.* Which of these angles would you expect, in general, to change with time?

2.15. The inclination of an orbit is 67°. What is the greatest latitude, north and south, reached by the subsatellite point? Is this orbit retrograde or prograde?

2.16. Describe briefly the main effects of the earth's equatorial bulge on a satellite orbit. Given that a satellite is in a circular equatorial orbit for which the semimajor axis is equal to 42,165 km, calculate (*a*) the mean motion, (*b*) the rate of regression of the nodes, and (*c*) the rate of rotation of argument of perigee.

2.17. A satellite in polar orbit has a perigee height of 600 km and an apogee height of 1200 km. Calculate (*a*) the mean motion, (*b*) the rate of regression of the nodes, and (c) the rate of rotation of the line of apsides. The mean radius of the earth may be assumed equal to 6371 km.

2.18. What is the fundamental unit of universal coordinated time? Express the following times in (*a*) days and (*b*) degrees: 0 h, 5 min, 24 s; 6 h, 35 min, 20 s; your present time.

2.19. Determine the Julian days for the following dates and times: midnight March 10, 1999; noon, February 23, 2000; 16:30 h, March 1, 2003; 3 P.M., July 4, 2010.

2.20. Find, for the times and dates given in Prob. 2.19, (*a*) T in Julian centuries and (*b*) the corresponding GST in degrees.

2.21. Find the month, day, and UT for the following epochs: (*a*) day 3.00, year 1999; (*b*) day 186.125, year 2000; (*c*) day 300.12157650, year 2001; (*d*) day 3.29441845, year 2004; (*e*) day 31.1015, year 2010.

2.22. Find the GST corresponding to the epochs given in Prob. 2.21.

2.23. The Molnya 3-(25) satellite has the following parameters specified: perigee height 462 km; apogee height 40,850 km; period 736 min; inclination 62.8°. Using an average value of 6371 km for the earth's radius, calculate (*a*) the semimajor axis and (*b*) the eccentricity. (*c*) Calculate the nominal mean motion n_0. (*d*) Calculate the mean motion. (*e*) Using the calculated value for a, calculate the anomalistic period and compare with the specified value. Calculate (*f*) the rate of regression of the nodes, and (*g*) the rate of rotation of the line of apsides.

2.24. Repeat the calculations in Prob. 2.23 for an inclination of 63.435°.

2.25. Determine the orbital condition necessary for the argument of perigee to remain stationary in the orbital plane. The orbit for a satellite under this condition has an eccentricity of 0.001 and a semimajor axis of 27,000 km. At a given epoch the perigee is exactly on the line of Aries. Determine the satellite position relative to this line after a period of 30 days from epoch.

2.26. For a given orbit, K as defined by Eq. (2.11) is equal to 0.112 rev/day. Determine the value of inclination required to make the orbit sun synchronous.

2.27. A satellite has an inclination of 90° and an eccentricity of 0.1. At epoch, which corresponds to time of perigee passage, the perigee height is 2643.24 km directly over the north pole. Determine (*a*) the satellite mean motion. For 1 day after epoch determine (*b*) the true anomaly, (*c*) the magnitude of the radius vector to the satellite, and (*d*) the latitude of the subsatellite point.

2.28. The following elements apply to a satellite in inclined orbit: $\Omega_0 = 0°$; $\omega_0 = 90°$; $M_0 = 309°$; $i = 63°$; $e = 0.01$; $a = 7130$ km. An earth station is situated at 45°N, 80°W, and at zero height above sea level. Assuming a perfectly spherical earth of uniform mass and radius 6371 km, and given that epoch corresponds to a GST of 116°, determine at epoch the orbital radius vector in the (*a*) **PQW** frame; (*b*) **IJK** frame; (*c*) the position vector of the earth station in the **IJK** frame; (*d*) the range vector in the **IJK** frame; (*e*) the range vector in the **SEZ** frame; and (*f*) the earth station look angles.

2.29. A satellite moves in an inclined elliptical orbit, the inclination being 63.45°. State with explanation the maximum northern and southern latitudes reached by the subsatellite point. The nominal mean motion of the satellite is 14 rev/day, and at epoch the subsatellite point is on the ascending node at 100°W. Calculate the longitude of the subsatellite point 1 day after epoch. The eccentricity is 0.01.

2.30. A "no name" satellite has the following parameters specified: perigee height 197 km; apogee height 340 km; period 88.2 min; inclination 64.6°. Using an average value of 6371 km for the earth's radius, calculate (*a*) the semimajor axis and (*b*) the eccentricity. (*c*) Calculate the nominal mean motion n_0. (*d*) Calculate the mean motion. (*e*) Using the calculated value for a, calculate the anomalistic period and compare with the specified value. Calculate (*f*) the rate of regression of the nodes, and (*g*) the rate of rotation of the line of apsides.

2.31. Given that $\Omega_0 = 250°$, $\omega_0 = 85°$, and $M_0 = 30°$ for the satellite in Prob. 2.30, calculate, for 65 min after epoch ($t_0 = 0$) the new values of Ω, ω, and M. Find also the true anomaly and radius.

2.32. From the NASA bulletin given in App. C, determine the date and the semimajor axis.

2.33. Determine, for the satellite listed in the NASA bulletin of App. C, the rate of regression of the nodes, the rate of change of the argument of perigee, and the nominal mean motion n_0.

2.34. From the NASA bulletin in App. C, verify that the orbital elements specified are for a nominal S–N equator crossing.

2.35. A satellite in exactly polar orbit has a slight eccentricity (just sufficient to establish the idea of a perigee). The anomalistic period is 110 min. Assuming that the mean motion is $n = n_0$ calculate the semimajor axis. Given that at epoch the perigee is exactly over the north pole, determine the position of the perigee relative to the north pole after one anomalistic period and the time taken for the satellite to make one complete revolution relative to the north pole.

2.36. A satellite is in an exactly polar orbit with apogee height 7000 km and perigee height 600 km. Assuming a spherical earth of uniform mass and radius 6371 km, calculate (*a*) the semimajor axis, (*b*) the eccentricity, and (*c*) the orbital period. (*d*) At a certain time the satellite is observed ascending directly overhead from an earth station on latitude 49°N. Give that the argument of perigee is 295° calculate the true anomaly at the time of observation.

2.37. The 2-line elements for satellite NOAA 18 are as follows:

NOAA 18
1 28654U 05018A 05154.51654998-.00000093 00000-0-28161-4 0 189
2 28654 98.7443 101.8853 0013815 210.8695 149.1647 14.10848892 1982

Determine the approximate values of (*a*) the semimajor axis, and (*b*) the latitude of the subsatellite point at epoch.

2.38. Using the 2-line elements given in Prob. 2.37, determine the longitude, of the subsatellite point and the LST at epoch.

2.39. Equation 2.34, gives the GST in degrees as

$$\text{GST} = 99.9610^\circ + 36000.7689^\circ \times T + 0.0004^\circ \times T^2 + \text{UT}^\circ$$

where T is the number of Julian centuries that have elapsed since noon, January 0, 1900. The GST equation is derived from (Wertz 1984) GST $= \alpha_s - 180^\circ + \text{UT}^\circ$ where α_s is the right ascension of the mean sun. Determine the right ascension of the mean sun for noon on June 5, 2005.

2.40. Assuming that the orbits detailed in Table 2.5 are circular, and using Eq. (2.2) to find the semimajor axis, calculate the regression of the nodes for these orbits.

2.41. Determine the standard zone time in the following zones, for 12 noon GMT: (*a*) 285°E, (*b*) 255°E, (*c*) 45°E, (*d*) 120°E.

2.42. Determine the GMT for the following local times and locations: (*a*) 7 A.M. Los Angeles, USA; (*b*) 1 P.M. Toronto, Canada; (*c*) 12 noon Baghdad, Iraq; (*d*) 3 P.M. Tehran, Iran.

References

ADC USAF. 1980. *Model for Propagation of NORAD Element Sets*. Spacetrack Report No. 3. Aerospace Defense Command, U.S. Air Force, December.

ADC USAF. 1983. *An Analysis of the Use of Empirical Atmospheric Density Models in Orbital Mechanics*. Spacetrack Report No. 3. Aerospace Defense Command, U.S. Air Force, February.

Arons, A. B. 1965. *Development of Concepts of Physics*. Addison-Wesley, Reading, MA.

Bate, R. R., D. D. Mueller, and J. E. White. 1971. *Fundamentals of Astrodynamics*. Dover, New York.

Celestrak, at http://celestrak.com/NORAD/elements/noaa.txt

Duffett-Smith, P. 1986. *Practical Astronomy with Your Calculator.* Cambridge University Press, New York.

Schwalb, A. 1982a. The TIROS-N/NOAA-G Satellite Series. NOAA Technical Memorandum NESS 95, Washington, DC.

Schwalb, A. 1982b. Modified Version of the TIROS-N/NOAA A-G Satellite Series (NOAA E-J): Advanced TIROS N (ATN). NOAA Technical Memorandum NESS 116, Washington, DC.

Thompson, Morris M. (editor-in-chief). 1966. *Manual of Photogrammetry*, 3d ed., Vol. 1. American Society of Photogrammetry, New York.

Wertz, J. R. (ed.). 1984. *Spacecraft Attitude Determination and Control.* D. Reidel, Holland.

Chapter

3

The Geostationary Orbit

3.1 Introduction

A satellite in a geostationary orbit appears to be stationary with respect to the earth, hence the name *geostationary*. Three conditions are required for an orbit to be geostationary:

1. The satellite must travel eastward at the same rotational speed as the earth.
2. The orbit must be circular.
3. The inclination of the orbit must be zero.

The first condition is obvious. If the satellite is to appear stationary, it must rotate at the same speed as the earth, which is constant. The second condition follows from this and from Kepler's second law (Sec. 2.3). Constant speed means that equal areas must be swept out in equal times, and this can only occur with a circular orbit (see Fig. 2.2). The third condition, that the inclination must be zero, follows from the fact that any inclination would have the satellite moving north and south, (see Sec. 2.5 and Fig. 2.3), and hence it would not be geostationary. Movement north and south can be avoided only with zero inclination, which means that the orbit lies in the earth's equatorial plane.

Kepler's third law may be used to find the radius of the orbit (for a circular orbit, the semimajor axis is equal to the radius). Denoting the radius by a_{GSO}, then from Eqs. (2.2) and (2.4),

$$a_{\mathrm{GSO}} = \left(\frac{\mu P}{4\pi^2}\right)^{1/3} \tag{3.1}$$

The period P for the geostationary is 23 h, 56 min, 4 s mean solar time (ordinary clock time). This is the time taken for the earth to complete one revolution about its N–S axis, measured relative to the fixed stars (see Sec. 2.9.4). Substituting this value along with the value for μ given by Eq. (2.3) results in

$$a_{\text{GSO}} = 42164 \text{ km} \tag{3.2}$$

The equatorial radius of the earth, to the nearest kilometer, is

$$a_E = 6378 \text{ km} \tag{3.3}$$

and hence the geostationary height is

$$\begin{aligned} h_{\text{GSO}} &= a_{\text{GSO}} - a_E \\ &= 42164 - 6378 \\ &= 35786 \text{ km} \end{aligned} \tag{3.4}$$

This value is often rounded up to 36,000 km for approximate calculations.

In practice, a precise geostationary orbit cannot be attained because of disturbance forces in space and the effects of the earth's equatorial bulge. The gravitational fields of the sun and the moon produce a shift of about 0.85°/year in inclination. Also, the earth's *equatorial ellipticity* causes the satellite to drift eastward along the orbit. In practice, station-keeping maneuvers have to be performed periodically to correct for these shifts, as described in Sec. 7.4.

An important point to grasp is that there is only one geostationary orbit because there is only one value of a that satisfies Eq. (2.3) for a periodic time of 23 h, 56 min, 4 s. Communications authorities throughout the world regard the geostationary orbit as a natural resource, and its use is carefully regulated through national and international agreements.

3.2 Antenna Look Angles

The *look angles* for the ground station antenna are the azimuth and elevation angles required at the antenna so that it points directly at the satellite. In Sec. 2.9.8 the look angles were determined in the general case of an elliptical orbit, and there the angles had to change in order to track the satellite. With the geostationary orbit, the situation is much simpler because the satellite is stationary with respect to the earth. Although in general no tracking should be necessary, with the large earth stations used for commercial communications, the antenna beamwidth is very narrow (see Chap. 6), and a tracking mechanism is required

to compensate for the movement of the satellite about the nominal geostationary position. With the types of antennas used for home reception, the antenna beamwidth is quite broad, and no tracking is necessary. This allows the antenna to be fixed in position, as evidenced by the small antennas used for reception of satellite TV that can be seen fixed to the sides of homes.

The three pieces of information that are needed to determine the look angles for the geostationary orbit are

1. The earth-station latitude, denoted here by λ_E
2. The earth-station longitude, denoted here by ϕ_E
3. The longitude of the subsatellite point, denoted here by ϕ_{SS} (often this is just referred to as the satellite longitude)

As in Chap. 2, latitudes north will be taken as positive angles, and latitudes south, as negative angles. Longitudes east of the Greenwich meridian will be taken as positive angles, and longitudes west, as negative angles. For example, if a latitude of 40°S is specified, this will be taken as $-40°$, and if a longitude of 35°W is specified, this will be taken as $-35°$.

In Chap. 2, when calculating the look angles for *low-earth-orbit* (LEO) satellites, it was necessary to take into account the variation in earth's radius. With the geostationary orbit, this variation has negligible effect on the look angles, and the average radius of the earth will be used. Denoting this by R:

$$R = 6371 \text{ km} \tag{3.5}$$

The geometry involving these quantities is shown in Fig. 3.1. Here, ES denotes the position of the earth station, SS the subsatellite point, S the satellite, and d is the range from the earth station to the satellite. The angle σ is an angle to be determined.

There are two types of triangles involved in the geometry of Fig. 3.1, the spherical triangle shown in heavy outline in Fig. 3.2*a* and the plane triangle of Fig. 3.2*b*. Considering first the spherical triangle, the sides are all arcs of great circles, and these sides are defined by the angles subtended by them at the center of the earth. Side a is the angle between the radius to the north pole and the radius to the subsatellite point, and it is seen that $a = 90°$. A spherical triangle in which one side is 90° is called a *quadrantal triangle*. Angle b is the angle between the radius to the earth station and the radius to the subsatellite point. Angle c is the angle between the radius to the earth station and the radius to the north pole. From Fig. 3.2*a* it is seen that $c = 90° - \lambda_E$.

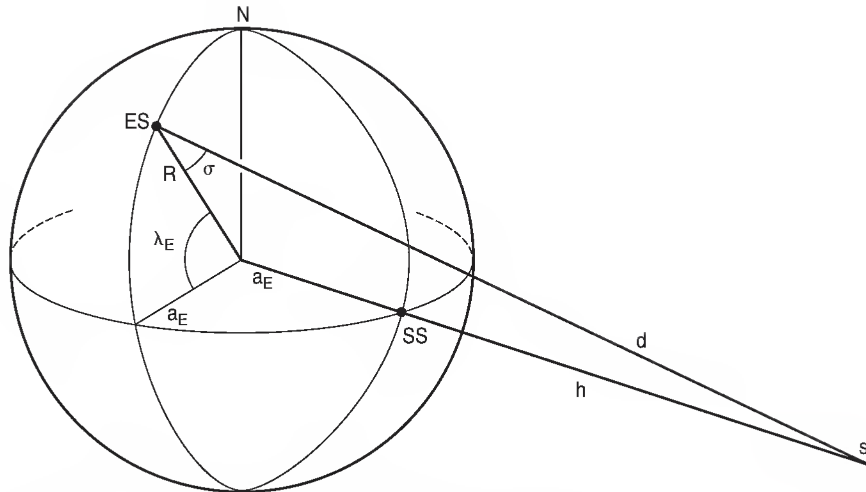
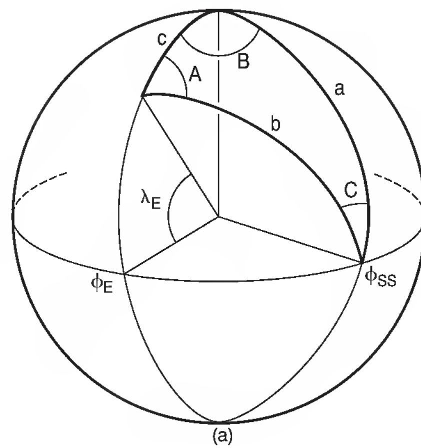


Figure 3.1 The geometry used in determining the look angles for a geostationary satellite.

There are six angles in all defining the spherical triangle. The three angles *A*, *B*, and *C* are the angles between the planes. Angle *A* is the angle between the plane containing *c* and the plane containing *b*. Angle *B* is the angle between the plane containing *c* and the plane containing *a*. From



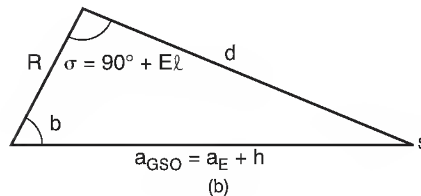


Figure 3.2 (*a*) The spherical geometry related to Fig. 3.1. (*b*) The plane triangle obtained from Fig. 3.1.

Fig. 3.2*a*, $B = \phi_E - \phi_{SS}$. It will be shown shortly that the maximum value of B is 81.3°. Angle C is the angle between the plane containing b and the plane containing a.

To summarize to this point, the information known about the spherical triangle is

$$a = 90° \tag{3.6}$$

$$c = 90° - \lambda_E \tag{3.7}$$

$$B = \phi_E - \phi_{SS} \tag{3.8}$$

Note that when the earth station is west of the subsatellite point, B is negative, and when east, B is positive. When the earth-station latitude is north, c is less than 90°, and when south, c is greater than 90°. Special rules, known as *Napier's rules*, are used to solve the spherical triangle (see Wertz, 1984), and these have been modified here to take into account the signed angles B and λ_E. Only the result will be stated here. Napier's rules gives angle b as

$$b = \arccos(\cos B \ \cos \lambda_E) \tag{3.9}$$

and angle A as

$$A = \arcsin\left(\frac{\sin|B|}{\sin b}\right) \tag{3.10}$$

Two values will satisfy Eq. (3.10), A and $180° - A$, and these must be determined by inspection. These are shown in Fig. 3.3. In Fig. 3.3*a*, angle A is acute (less than 90°), and the azimuth angle is $A_z = A$. In Fig. 3.3*b*, angle A is acute, and the azimuth is, by inspection, $A_z = 360° - A$. In Fig. 3.3*c*, angle A_c is obtuse and is given by $A_c = 180° - A$, where A is the acute value obtained from Eq. (3.10). Again, by inspection, $A_z = A_c - 180° - A$. In Fig. 3.3*d*, angle A_d is obtuse and is given by $180° - A$, where A is the acute value obtained from Eq. (3.10). By inspection, $A_z = 360° - A_d = 180° + A$. In all cases, A is the acute angle returned by Eq. (3.10). These conditions are summarized in Table 3.1.

Example 3.1 A geostationary satellite is located at 90°W. Calculate the azimuth angle for an earth-station antenna at latitude 35°N and longitude 100°W.

Solution The given quantities are
$\phi_{SS} = -90°$ $\quad \phi_E = -100°$ $\quad \lambda_E = 35°$

$$B = \phi_E - \phi_{SS}$$
$$= -10°$$

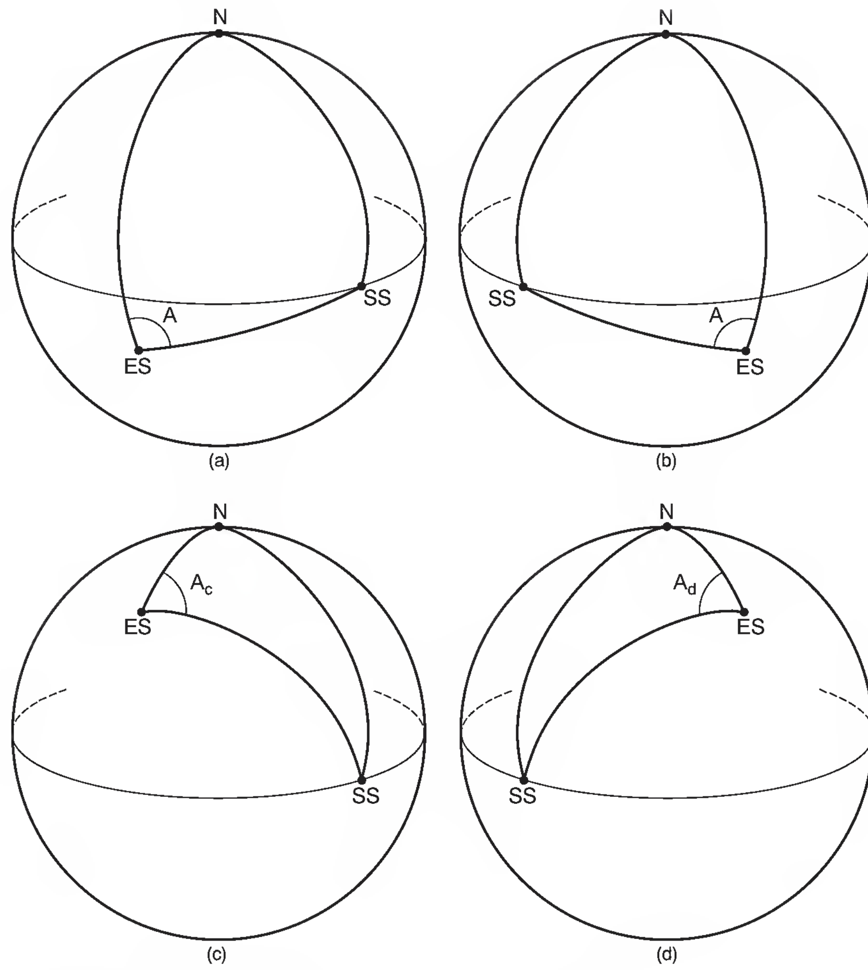


Figure 3.3 Azimuth angles related to angle A (see Table 3.1).

TABLE 3.1 Azimuth Angles A_z from Fig. 3.3

Fig. 3.3	λ_E	B	A_z, degrees
a	<0	<0	A
b	<0	>0	$360° - A$
c	>0	<0	$180° - A$
d	>0	>0	$180° + A$

From Eq. (3.8):

$$b = \arccos(\cos B \ \cos \lambda_E)$$
$$= 36.23°$$

From Eq. (3.9):

$$A = \arcsin\left(\frac{\sin|B|}{\sin b}\right)$$
$$= 17.1°$$

By inspection, $\lambda_E > 0$ and B < 0. Therefore, Fig. 3.3*c* applies, and

$$A_z = 180° - A$$
$$= \underline{\underline{162.9°}}$$

Applying the cosine rule for plane triangles to the triangle of Fig. 3.2*b* allows the range *d* to be found to a close approximation:

$$d = \sqrt{R^2 + a_{GSO}^2 - 2Ra_{GSO}\cos b} \tag{3.11}$$

Applying the sine rule for plane triangles to the triangle of Fig. 3.2*b* allows the angle of elevation to be found:

$$El = \arccos\left(\frac{a_{GSO}}{d}\sin b\right) \tag{3.12}$$

Example 3.2 Find the range and antenna elevation angle for the situation specified in Example 3.1.

Solution R = 6371 km; a_{GSO} = 42164 km, and from Example 3.1, b = 36.23°. Equation (3.11) gives:

$$d = \sqrt{6371^2 + 42164^2 - 2 \times 6371 \times 42164 \times \cos 36.23°}$$
$$\cong \underline{\underline{37215 \text{ km}}}$$

Equation (3.11) gives:

$$El = \arccos\left(\frac{42164}{37215}\sin 36.23°\right)$$
$$\cong \underline{\underline{48°}}$$

Figure 3.4 shows the look angles for Ku-band satellites as seen from Thunder Bay, Ontario, Canada.

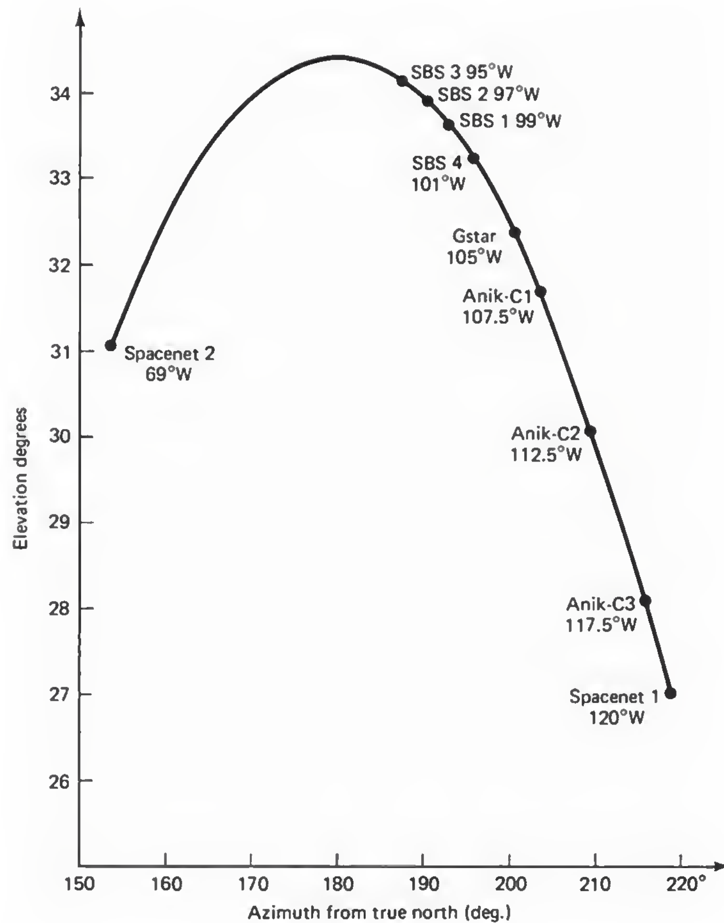


Figure 3.4 Azimuth-elevation angles for an earth-station location of 48.42°N, 89.26°W (Thunder Bay, Ontario). Ku-band satellites are shown.

The preceding results do not take into account the case when the earth station is on the equator. Obviously, when the earth station is directly under the satellite, the elevation is 90°, and the azimuth is irrelevant. When the subsatellite point is east of the equatorial earth station ($B < 0$), the azimuth is 90°, and when west ($B > 0$), the azimuth is 270°. Also, the range as determined by Eq. (3.11) is approximate, and where more accurate values are required, as, for example, where propagation times need to be known accurately, the range is determined by measurement.

For a typical home installation, practical adjustments will be made to align the antenna to a known satellite for maximum signal. Thus the look angles need not be determined with great precision but are calculated

to give the expected values for a satellite the longitude of which is close to the earth-station longitude. In some cases, especially with *direct broadcast satellites* (DBS), the home antenna is aligned to one particular cluster of satellites, as described in Chap. 16, and no further adjustments are necessary.

3.3 The Polar Mount Antenna

Where the home antenna has to be steerable, expense usually precludes the use of separate azimuth and elevation actuators. Instead, a single actuator is used which moves the antenna in a circular arc. This is known as a *polar mount antenna.* The antenna pointing can only be accurate for one satellite, and some pointing error must be accepted for satellites on either side of this. With the polar mount antenna, the dish is mounted on an axis termed the *polar axis* such that the antenna boresight is normal to this axis, as shown in Fig. 3.5*a*. The polar mount is aligned along a true north line, as shown in Fig. 3.5, with the boresight pointing due south. The angle between the polar mount and the local horizontal plane is set equal to the earth-station latitude λ_E; simple geometry shows that this makes the boresight lie parallel to the equatorial plane. Next, the dish is tilted at an angle δ relative to the polar mount until the boresight is pointing at a satellite position due south of the earth station. Note that there does not need to be an actual satellite at this position. (The angle of tilt is often referred to as the *declination*, which must not be confused with the magnetic declination used in correcting compass readings. The term *angle of tilt* will be used for δ in this text.)

The required angle of tilt is found as follows: From the geometry of Fig. 3.5*b*,

$$\delta = 90° - El_0 - \lambda_E \tag{3.13}$$

where El_0 is the angle of elevation required for the satellite position due south of the earth station. But for the due south situation, angle B in Eq. (3.8) is equal to zero; hence, from Eq. (3.9), $b = \lambda_E$. Hence, from Eq. (3.12), or Fig 3.5*c*.

$$\cos El_0 = \frac{a_{GSO}}{d} \sin \lambda_E \tag{3.14}$$

Combining Eqs. (3.13) and (3.14) gives the required angle of tilt as

$$\delta = 90° - \arccos\left(\frac{a_{GSO}}{d}\right) \sin \lambda_E - \lambda_E \tag{3.15}$$

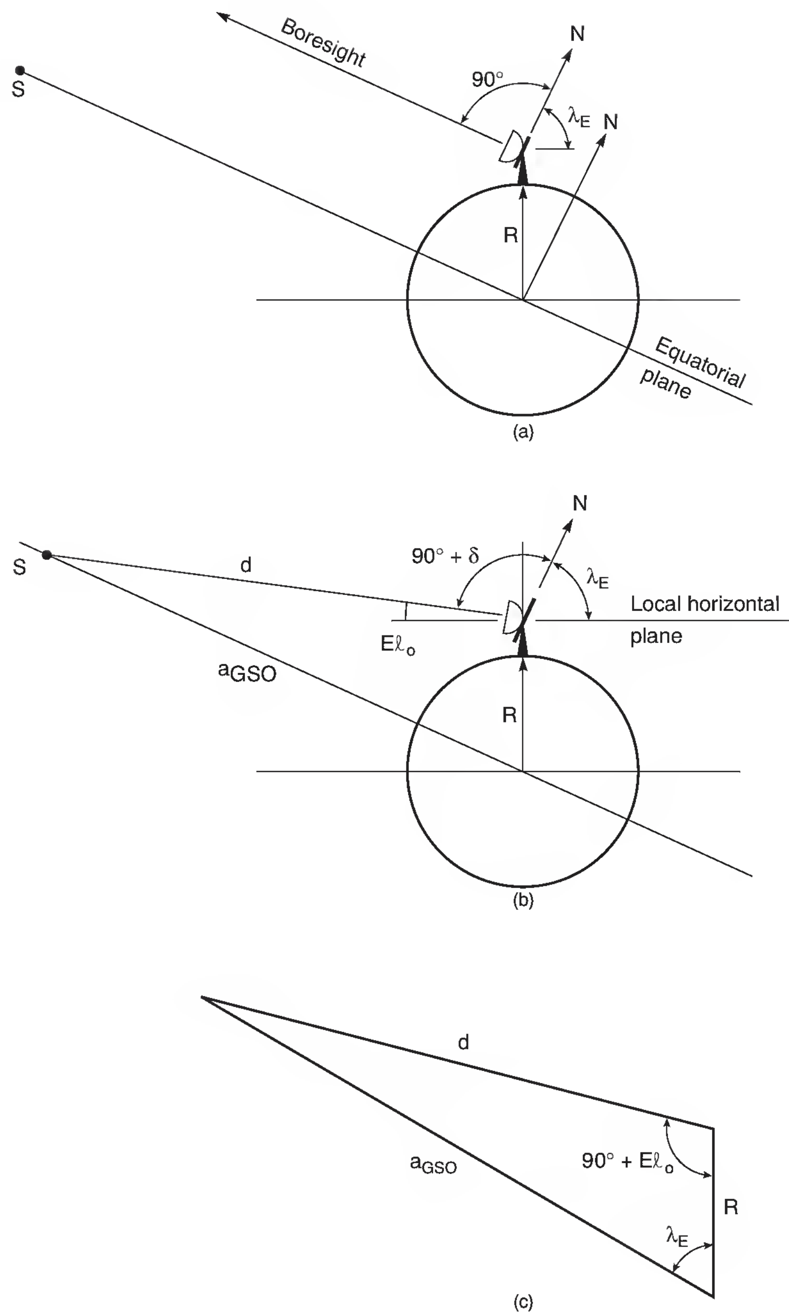


Figure 3.5 The polar mount antenna.

In the calculations leading to d, a spherical earth of mean radius 6371 km may be assumed and earth-station elevation may be ignored, as was done in the previous section. The value obtained for δ will be sufficiently accurate for initial alignment and fine adjustments can be made, if necessary. Calculation of the angle of tilt is illustrated in Example 3.3.

Example 3.3 Determine the angle of tilt required for a polar mount used with an earth station at latitude 49° north. Assume a spherical earth of mean radius 6371 km, and ignore earth-station altitude.

Solution Given data:

$\lambda_E = 49°$; $a_{GSO} = 42164$ km; $R = 6371$ km; $b = \lambda_E = 49°$.

Equation (3.11) gives:

$$d = \sqrt{6371^2 + 42164^2 - 2 \times 6371 \times 42164 \times \cos 49°}$$
$$\cong 38287 \text{ km}$$

From Eq. (3.12):

$$El = \arccos\left(\frac{42164}{38287} \sin 49°\right)$$
$$\cong 33.8°$$
$$\delta = 90° - 33.8° - 49°$$
$$\cong \underline{\underline{7°}}$$

3.4 Limits of Visibility

There will be east and west limits on the geostationary arc visible from any given earth station. The limits will be set by the geographic coordinates of the earth station and the antenna elevation. The lowest elevation in theory is zero, when the antenna is pointing along the horizontal. A quick estimate of the longitudinal limits can be made by considering an earth station at the equator, with the antenna pointing either west or east along the horizontal, as shown in Fig. 3.6. The limiting angle is given by

$$\theta = \arccos \frac{a_E}{a_{GSO}}$$
$$= \arccos \frac{6378}{42164}$$
$$= 81.3°$$

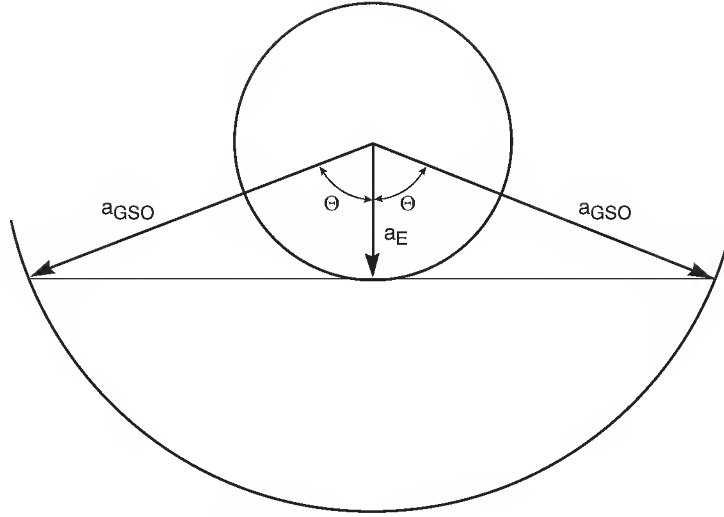


Figure 3.6 Illustrating the limits of visibility.

Thus, for this situation, an earth station could see satellites over a geostationary arc bounded by ±81.3° about the earth-station longitude.

In practice, to avoid reception of excessive noise from the earth, some finite minimum value of elevation is used, which will be denoted here by El_{min}. A typical value is 5°. The limits of visibility will also depend on the earth-station latitude. As in Fig. 3.2*b*, let S represent the angle subtended at the satellite when the angle $\sigma_{\text{min}} = 90° + El_{\text{min}}$. Applying the sine rule gives

$$S = \arcsin\left(\frac{R}{a_{\text{GSO}}}\sin\sigma_{\text{min}}\right) \tag{3.17}$$

A sufficiently accurate estimate is obtained by assuming a spherical earth of mean radius 6371 km as was done previously. Once angle S is known, angle b is found from

$$b = 180 - \sigma_{\text{min}} - S \tag{3.18}$$

From Eq. (3.9):

$$B = \arccos\left(\frac{\cos b}{\cos \lambda_E}\right) \tag{3.19}$$

Once angle B is found, the satellite longitude can be determined from Eq. (3.8). This is illustrated in Example 3.4.

Example 3.4 Determine the limits of visibility for an earth station situated at mean sea level, at latitude 48.42° north, and longitude 89.26 degrees west. Assume a minimum angle of elevation of 5°.

Solution Given data:
$\lambda_E = 48.42°$; $\phi_E = -89.26°$; $El_{\min} = 5°$; $a_{\mathrm{GSO}} = 42164$ km; $R = 6371$ km

$$\sigma_{\min} = 90° + El_{\min}$$

Equation (3.17) gives:

$$S = \arcsin\left(\frac{6371}{42164}\sin 95°\right)$$
$$= 8.66°$$

Equation (3.18) gives:

$$b = 180 - 95° - 8.66°$$
$$= 76.34°$$

Equation (3.19) gives:

$$B = \arccos\left(\frac{\cos 76.34°}{\cos 48.42°}\right)$$
$$= 69.15°$$

The satellite limit east of the earth station is at

$$\phi_E + B = \underline{\underline{-20°}} \text{ approx.}$$

and west of the earth station at

$$\phi_E - B = \underline{\underline{-158°}} \text{ approx.}$$

3.5 Near Geostationary Orbits

As mentioned in Sec. 2.8, there are a number of perturbing forces that cause an orbit to depart from the ideal keplerian orbit. For the geostationary case, the most important of these are the gravitational fields of the moon and the sun, and the nonspherical shape of the earth. Other significant forces are solar radiation pressure and reaction of the satellite itself to motor movement within the satellite. As a result, station-keeping maneuvers must be carried out to maintain the satellite within set limits of its nominal geostationary position. Station keeping is discussed in Sec. 7.4.

An exact geostationary orbit therefore is not attainable in practice, and the orbital parameters vary with time. The two-line orbital elements

are published at regular intervals, Fig. 3.7 showing typical values. The period for a geostationary satellite is 23 h, 56 min, 4 s, or 86,164 s. The reciprocal of this is 1.00273896 rev/day, which is about the value tabulated for most of the satellites in Fig. 3.7. Thus these satellites are *geosynchronous,* in that they rotate in synchronism with the rotation of the earth. However, they are not geostationary. The term *geosynchronous satellite* is used in many cases instead of *geostationary* to describe these near-geostationary satellites. It should be noted, however, that in general a geosynchronous satellite does not have to be near-geostationary, and there are a number of geosynchronous satellites that are in highly elliptical orbits with comparatively large inclinations (e.g., the Tundra satellites).

Although in principle the two-line elements could be used as described in Chap. 2 to determine orbital motion, the small inclination makes it difficult to locate the position of the ascending node, and the small eccentricity makes it difficult to locate the position of the perigee. However, because of the small inclination, the angles ω and Ω can be assumed to be in the same plane.

Referring to Fig. 2.9 it will be seen that with this assumption the subsatellite point will be $\Omega + \omega + v$ east of the line of Aries. The longitude of the subsatellite point (the satellite longitude) is the easterly

```
INTELSAT 901
1 26824U 01024A   05122.92515626 -.00000151  00000-0  10000-3 0  7388
2 26824   0.0158 338.7780 0004091  67.7508 129.4375  1.00270746 14318
INTELSAT 902
1 26900U 01039A   05126.99385197  .00000031  00000-0  10000-3 0  6260
2 26900   0.0156 300.5697 0002640 112.8823 231.2391  1.00271845 13528
INTELSAT 903
1 27403U 02016A   05125.03556931  .00000000  00000-0  10000-3 0  6249
2 27403   0.0362 171.6123 0002986 232.5077 157.1571  1.00265355 11412
INTELSAT 904
1 27380U 02007A   05125.62541657  .00000043  00000-0  00000+0 0  5361
2 27380   0.0202   0.0174 0003259  40.3723 108.3316  1.00272510 11761
INTELSAT 905
1 27438U 02027A   05125.03693822  .00000000  00000-0  10000-3 0  5812
2 27438   0.0205 164.2424 0002820 218.0675 189.4691  1.00265924 10746
INTELSAT 906
1 27513U 02041A   05126.63564565  .00000012  00000-0  00000+0 0  4817
2 27513   0.0111 324.7901 0003200  99.2828  93.4848  1.00272600  9803
INTELSAT 907
1 27683U 03007A   05124.32309516  .00000000  00000-0  10000-3 0  3108
2 27683   0.0206  13.5522 0009594  61.6856 235.7624  1.00266570  8131
INTELSAT 1002
1 28358U 04022A   05124.94126775 -.00000018  00000-0  00000+0 0  1527
2 28358   0.0079 311.0487 0000613  59.4312 190.2817  1.00271159  3289
```

Figure 3.7 Two-line elements for some geostationary satellites.

rotation from the Greenwich meridian. The *Greenwich sidereal time* (GST) gives the eastward position of the Greenwich meridian relative to the line of Aries, and hence the subsatellite point is at longitude

$$\phi_{SS} = \omega + \Omega + v - \text{GST} \tag{3.20}$$

and the mean longitude of the satellite is given by

$$\phi_{SSmean} = \omega + \Omega + M - \text{GST} \tag{3.21}$$

Equation (2.31) can be used to calculate the true anomaly, and because of the small eccentricity, this can be approximated as

$$v = M + 2e \sin M \tag{3.22}$$

The two-line elements for the Intelsat series, obtained from Celestrak at http://celestrak.com/NORAD/elements/intelsat.txt are shown in Fig. 3.7.

Example 3.5 Using the data given in Fig. 3.7, calculate the longitude for INTELSAT 10-02.

Solution From Fig. 3.7 the inclination is seen to be 0.0079°, which makes the orbit almost equatorial. Also the revolutions per day are 1.00271159, or approximately geosynchronous. Other values taken from Fig. 3.7 are:

epoch day = 124.94126775 days; year = 2005; $\Omega = 311.0487°$; $\omega = 59.4312°$; $M = 190.2817°$; $e = 0.0000613$

From Table 2.2 the Julian day for $\text{Jan}_{0.0}$ 2005 is $\text{JD}_{00} = 2453370.5$ days. The Julian day for epoch is JD = 2453370.5 + 124.94126775 = 2453495.44126775 days. The reference value is (see Eq. 2.20) $\text{JD}_{ref} = 2415020$ days. Hence T in Julian centuries is:

$$\begin{aligned} T &= \frac{\text{JD} - \text{JD}_{ref}}{36525} \\ &= \frac{38475.442}{36525} \\ &= 1.05340017 \end{aligned}$$

The decimal fraction of the epoch gives the UT as a fraction of a day, and in degrees this is:

$$\begin{aligned} \text{UT}° &= 0.94126775 \times 360° \\ &= 338.85637° \end{aligned}$$

Substituting these values in Eq. (2.34) gives, for the GST:

$$\begin{aligned} \text{GST} &= 99.9610° + 36000.7689° \times T + 0.0004° \times T^2 + \text{UT}° \\ &= 201.764° \text{ (mod } 360°) \end{aligned}$$

Equation (3.22) gives:

$$\nu = M + 2e \sin M$$
$$= 3.32104 \text{ rad} + 2 \times .0000613 \times \sin 190.2817°$$
$$= 190.28044°$$

Equation (3.20) then gives:

$$\phi_{SS} = \omega + \Omega + v - \text{GST}$$
$$= 59.4313° + 311.0487° + 190.2804° - 201.764°$$
$$= \underline{\underline{358.996°}}$$

and Eq. (3.21):

$$\phi_{SSmean} = \omega + \Omega + M - \text{GST}$$
$$= 59.4313° + 311.0487° + 190.2804° - 201.764°$$
$$= \underline{\underline{358.996°}}$$

From Table 1.3 the assigned spot for INTELSAT 10-02 is 359° east.

Modified inclination and eccentricity parameters can be derived from the specified values of inclination *i*, the eccentricity *e*, and the angles ω and Ω. Details of these will be found in Maral and Bousquet (1998).

3.6 Earth Eclipse of Satellite

If the earth's equatorial plane coincided with the plane of the earth's orbit around the sun (the ecliptic plane), geostationary satellites would be eclipsed by the earth once each day. As it is, the equatorial plane is tilted at an angle of 23.4° to the ecliptic plane, and this keeps the satellite in full view of the sun for most days of the year, as illustrated by position *A* in Fig. 3.8. Around the spring and autumnal equinoxes, when the sun is crossing the equator, the satellite does pass into the earth's shadow at certain periods, these being periods of eclipse as illustrated in Fig. 3.8. The spring equinox is the first day of spring, and the autumnal equinox is the first day of autumn.

Eclipses begin 23 days before equinox and end 23 days after equinox. The eclipse lasts about 10 min at the beginning and end of the eclipse period and increases to a maximum duration of about 72 min at full eclipse (Spilker, 1977). During an eclipse, the solar cells do not function, and operating power must be supplied from batteries. This is discussed further in Sec. 7.2, and Fig. 7.3 shows eclipse time as a function of days of the year.

Where the satellite longitude is east of the earth station, the satellite enters eclipse during daylight (and early evening) hours for the earth station, as illustrated in Fig. 3.9. This can be undesirable if the satellite

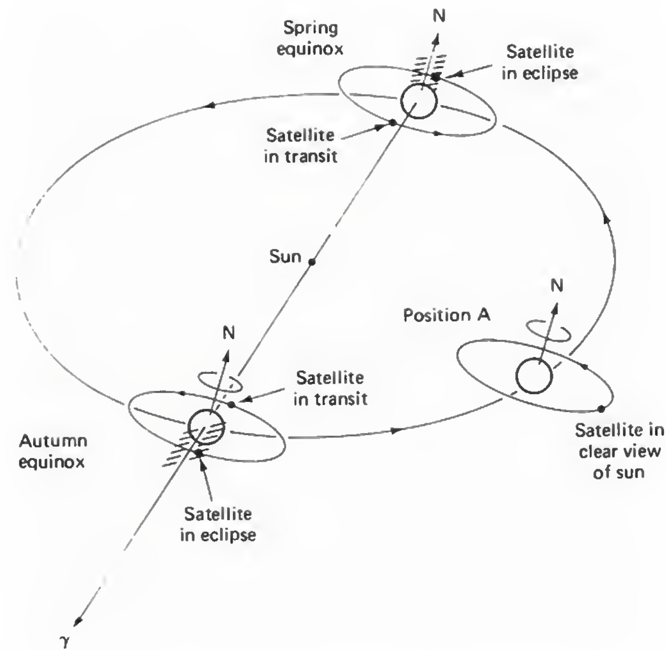


Figure 3.8 Showing satellite eclipse and satellite sun transit around spring and autumn equinoxes.

has to operate on reduced battery power. Where the satellite longitude is west of the earth station, eclipse does not occur until the earth station is in darkness, (or early morning) when usage is likely to be low. Thus satellite longitudes which are west, rather than east, of the earth station are more desirable.

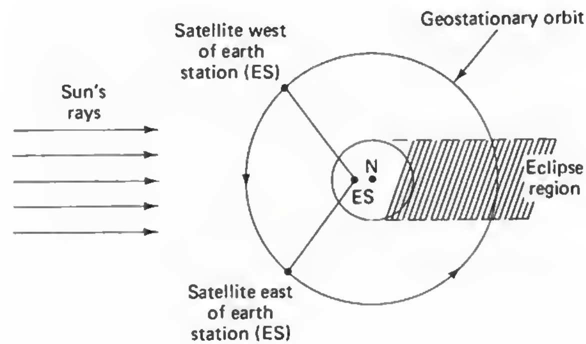


Figure 3.9 A satellite east of the earth station enters eclipse during daylight and early evening (busy) hours at the earth station. A satellite west of the earth station enters eclipse during night and early morning (nonbusy) hours.

3.7 Sun Transit Outage

Another event which must be allowed for during the equinoxes is the transit of the satellite between earth and sun (see Fig. 3.8), such that the sun comes within the beamwidth of the earth-station antenna. When this happens, the sun appears as an extremely noisy source which completely blanks out the signal from the satellite. This effect is termed *sun transit outage,* and it lasts for short periods—each day for about 6 days around the equinoxes. The occurrence and duration of the sun transit outage depends on the latitude of the earth station, a maximum outage time of 10 min being typical.

3.8 Launching Orbits

Satellites may be *directly injected* into low-altitude orbits, up to about 200 km altitude, from a launch vehicle. Launch vehicles may be classified as *expendable* or *reusable.* Typical of the expendable launchers are the U.S. Atlas-Centaur and Delta rockets and the European Space Agency Ariane rocket. Japan, China, and Russia all have their own expendable launch vehicles, and one may expect to see competition for commercial launches among the countries which have these facilities.

Until the tragic mishap with the Space Shuttle in 1986, this was to be the primary transportation system for the United States. As a reusable launch vehicle, the shuttle, also referred to as the *Space Transportation System* (STS), was planned to eventually replace expendable launch vehicles for the United States (Mahon and Wild, 1984).

Where an orbital altitude greater than about 200 km is required, it is not economical in terms of launch vehicle power to perform direct injection, and the satellite must be placed into transfer orbit between the initial LEO and the final high-altitude orbit. In most cases, the transfer orbit is selected to minimize the energy required for transfer, and such an orbit is known as a *Hohmann transfer* orbit. The time required for transfer is longer for this orbit than all other possible transfer obits.

Assume for the moment that all orbits are in the same plane and that transfer is required between two circular orbits, as illustrated in Fig. 3.10. The Hohmann elliptical orbit is seen to be tangent to the low-altitude orbit at perigee and to the high-altitude orbit at apogee. At the perigee, in the case of rocket launch, the rocket injects the satellite with the required thrust into the transfer orbit. With the STS, the satellite must carry a perigee kick motor which imparts the required thrust at perigee. Details of the expendable vehicle launch are shown in Fig. 3.11, and of the STS launch in Fig. 3.12. At apogee, the *apogee kick motor* (AKM) changes the velocity of the satellite to place it into a circular orbit

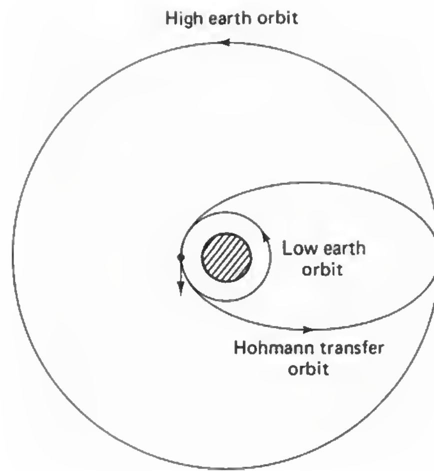


Figure 3.10 Hohmann transfer orbit.

in the same plane. As shown in Fig. 3.11, it takes 1 to 2 months for the satellite to be fully operational (although not shown in Fig. 3.12, the same conditions apply). Throughout the launch and acquisition phases, a network of ground stations, spread across the earth, is required to perform the *tracking, telemetry, and command* (TT&C) functions.

Velocity changes in the same plane change the geometry of the orbit but not its inclination. In order to change the inclination, a velocity change is required normal to the orbital plane. Changes in inclination can be made at either one of the nodes, without affecting the other orbital parameters. Since energy must be expended to make any orbital changes, a geostationary satellite should be launched initially with as low an orbital inclination as possible. It will be shown shortly that the smallest inclination obtainable at initial launch is equal to the latitude of the launch site. Thus the farther away from the equator a launch site is, the less useful it is, since the satellite has to carry extra fuel to effect a change in inclination. Russia does not have launch sites south of 45°N, which makes the launching of geostationary satellites a much more expensive operation for Russia than for other countries which have launch sites closer to the equator.

Prograde (direct) orbits (Fig. 2.4) have an easterly component of velocity, so prograde launches gain from the earth's rotational velocity. For a given launcher size, a significantly larger payload can be launched in an easterly direction than is possible with a retrograde (westerly) launch. In particular, easterly launches are used for the initial launch into the geostationary orbit.

The relationship between inclination, latitude, and azimuth may be seen as follows [this analysis is based on that given in Bate et al. (1971)]. Figure 3.13a shows the geometry at the launch site A at latitude λ (the

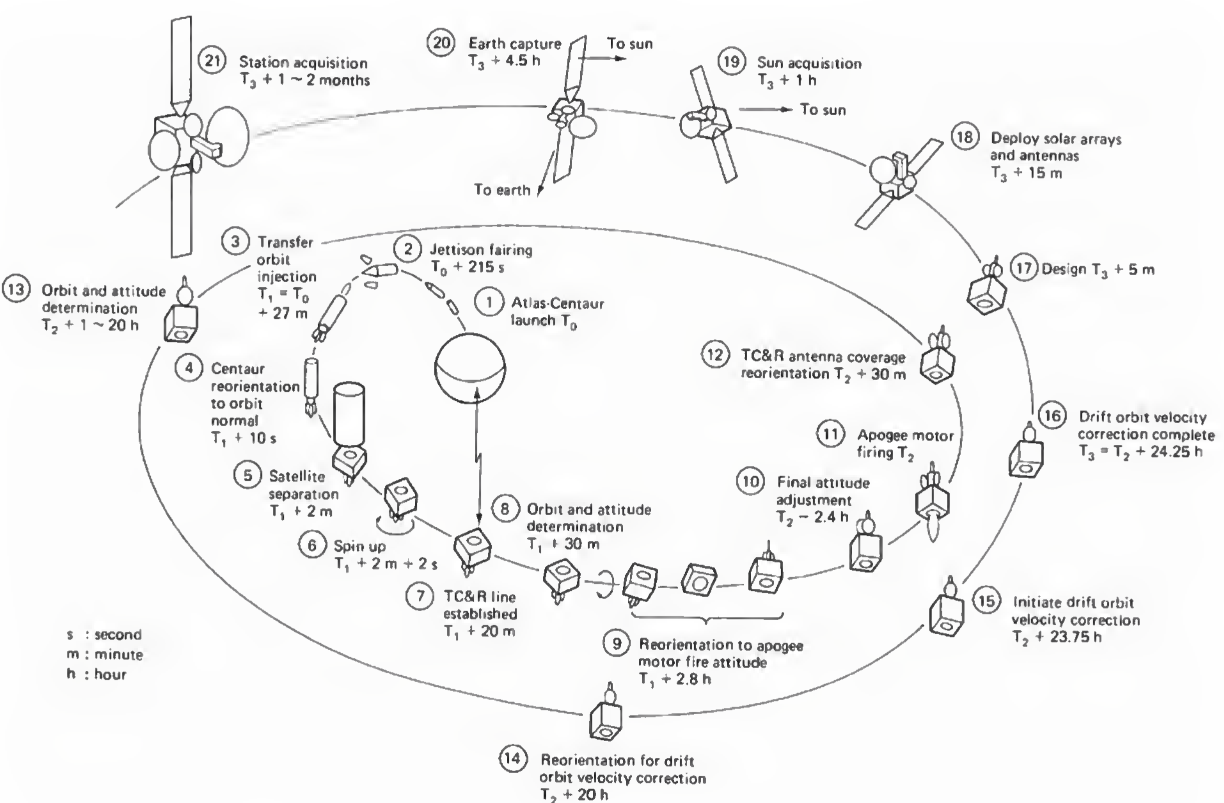


Figure 3.11 From launch to station of INTELSAT V (by Atlas-Centaur). (*From Satellite Communications Technology, edited by K. Miya, 1981; courtesy of KDD Engineering & Consulting, Inc., Tokyo.*)

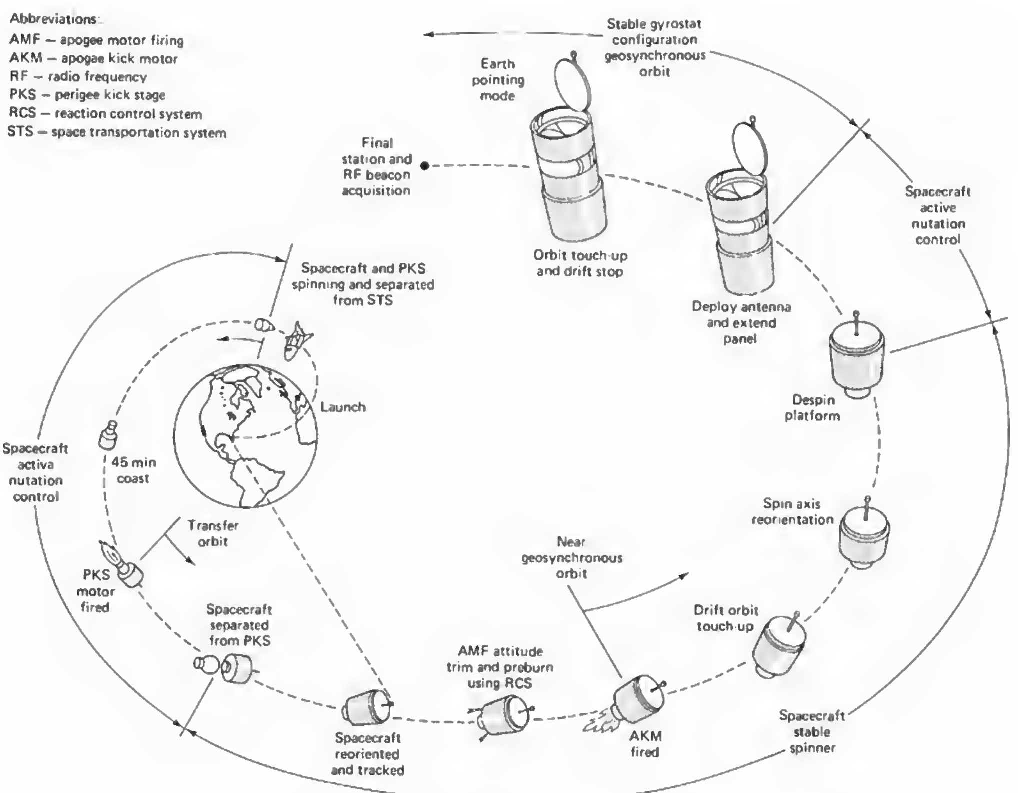


Figure 3.12 STS-7/Anik C2 mission scenario. (*From Anik C2 Launch Handbook; courtesy of Telesat, Canada.*)

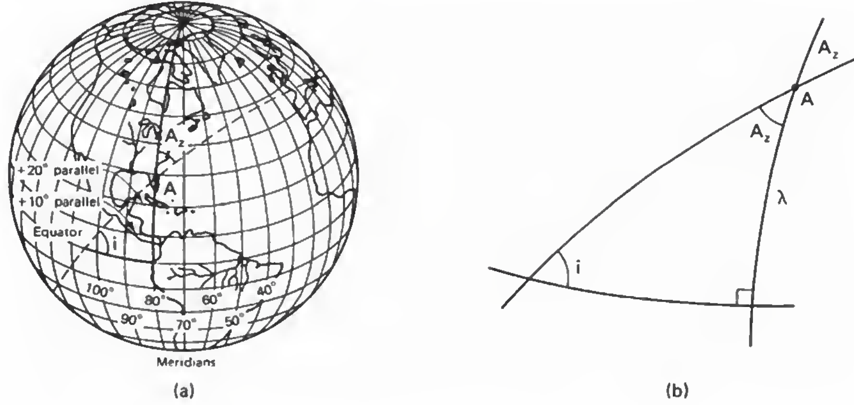


Figure 3.13 (*a*) Launch site *A*, showing launch azimuth A_z; (*b*) enlarged version of the spherical triangle shown in (*a*). λ is the latitude of the launch site.

slight difference between geodetic and geocentric latitudes may be ignored here). The dotted line shows the satellite earth track, the satellite having been launched at some azimuth angle A_z. Angle i is the resulting inclination.

The spherical triangle of interest is shown in more detail in Fig. 3.13*b*. This is a right spherical triangle, and Napier's rule for this gives

$$\cos i = \cos\lambda \sin A_z \tag{3.23}$$

For a prograde orbit (see Fig. 2.4 and Sec. 2.5), $0 \le i \le 90°$, and hence cos i is positive. Also, $-90° \le \lambda \le 90°$, and hence cos λ is also positive. It follows therefore from Eq. (3.23) that $0 \le A_z \le 180°$, or the launch azimuth must be easterly in order to obtain a prograde orbit, confirming what was already known.

For a fixed λ, Eq. (3.23) also shows that to minimize the inclination i, cos i should be a maximum, which requires sin A_z to be maximum, or $A_z = 90°$. Equation (3.23) shows that under these conditions

$$\cos i_{\min} = \cos\lambda \tag{3.24}$$

or

$$i_{\min} = \lambda \tag{3.25}$$

Thus the *lowest* inclination possible on initial launch is equal to the latitude of the launch site. This result confirms the converse statement made in Sec. 2.5 under *inclination* that the greatest latitude north or south is equal to the inclination. From Cape Kennedy the smallest initial inclination which can be achieved for easterly launches is approximately 28°.

3.9 Problems

3.1. Explain what is meant by the geostationary orbit. How do the geostationary orbit and a geosynchronous orbit differ?

3.2. (a) Explain why there is only one geostationary orbit. (b) Show that the range d from an earth station to a geostationary satellite is given by

$$d = \sqrt{(R \sin El)^2 + h(2R + h)} - R \sin El,$$

where R is the earth's radius (assumed spherical), h is the height of the geostationary orbit above the equator, and El is the elevation angle of the earth station antenna.

3.3. Determine the latitude and longitude of the farthest north earth station which can link with any given geostationary satellite. The longitude should be given relative to the satellite longitude, and a minimum elevation angle of 5° should be assumed for the earth station antenna. A spherical earth of mean radius 6371 km may be assumed.

3.4. An earth station at latitude 30°S is in communication with an earth station on the same longitude at 30°N, through a geostationary satellite. The satellite longitude is 20° east of the earth stations. Calculate the antenna-look angles for each earth station and the round-trip time, assuming this consists of propagation delay only.

3.5. Determine the maximum possible longitudinal separation which can exist between a geostationary satellite and an earth station while maintaining line-of-sight communications, assuming the minimum angle of elevation of the earth station antenna is 5°. State also the latitude of the earth station.

3.6. An earth station is located at latitude 35°N and longitude 100°W. Calculate the antenna-look angles for a satellite at 67°W.

3.7. An earth station is located at latitude 12°S and longitude 52°W. Calculate the antenna-look angles for a satellite at 70°W.

3.8. An earth station is located at latitude 35°N and longitude 65°E. Calculate the antenna-look angles for a satellite at 19°E.

3.9. An earth station is located at latitude 30°S and longitude 130°E. Calculate the antenna-look angles for a satellite at 156°E.

3.10. Calculate for your home location the look angles required to receive from the satellite (*a*) immediately east and (*b*) immediately west of your longitude.

3.11. CONUS is the acronym used for the 48 contiguous states. Allowing for a 5° elevation angle at earth stations, verify that the geostationary arc required to cover CONUS is 55° to 136°W.

3.12. Referring to Prob. 3.11, verify that the geostationary arc required for CONUS plus Hawaii is 85° to 136° W and for CONUS plus Alaska is 115° to 136°W.

3.13. By taking the Mississippi River as the dividing line between east and west, verify that the western region of the United States would be covered by satellites in the geostationary arc from 136° to 163°W and the eastern region by 25° to 55°W. Assume a 5° angle of elevation.

3.14. (*a*) An earth station is located at latitude 35°N. Assuming a polar mount antenna is used, calculate the angle of tilt. (*b*) Would the result apply to polar mounts used at the earth stations specified in Probs. 3.6 and 3.8?

3.15. Repeat Prob. 3.14 (*a*) for an earth station located at latitude 12°S. Would the result apply to a polar mount used at the earth station specified in Prob. 3.7?

3.16. Repeat Prob. 3.14 (*a*) for an earth station located at latitude 30°S. Would the result apply to a polar mount used at the earth station specified in Prob. 3.9?

3.17. Calculate the angle of tilt required for a polar mount antenna used at your home location.

3.18. The borders of a certain country can be roughly represented by a triangle with coordinates 39°E, 33.5°N; 43.5°E, 37.5°N; 48.5°E, 30°N. If a geostationary satellite has to be visible from *any point* in the country, determine the limits of visibility (i.e., the limiting longitudinal positions for a satellite on the geostationary arc). Assume a minimum angle of elevation for the earth station antenna of 5°, and show which geographic location fixes which limit.

3.19. Explain what is meant by the *earth eclipse* of an earth-orbiting satellite. Why is it preferable to operate with a satellite positioned west, rather than east, of earth station longitude?

3.20. Explain briefly what is meant by *sun transit outage*.

3.21. Using the data given in Fig. 3.7, calculate the longitude for INTELSAT 904.

3.22. Calculate the semimajor axis for INTELSAT 901.

3.23. Calculate the apogee and perigee heights for INTELSAT 906.

3.24. Calculate the rate of regression of the nodes and the rate of rotation of the line of apsides for INTELSAT 907.

References

Bate, R. R., D. D. Mueller, and J. E. White. 1971. *Fundamentals of Astrodynamics.* Dover, New York.

Celestrak, at http://celestrak.com/NORAD/elements/intelsat.txt

Mahon, J., and J. Wild. 1984. "Commercial Launch Vehicles and Upper Stages." *Space Commun. Broadcast.*, Vol. 2, pp. 339–362.

Maral, G., and M. Bousquet. 1998. *Satellite Communications Systems.* Wiley, New York.

Spilker, J. J. 1977. *Digital Communications by Satellite.* Prentice-Hall, Englewood Cliffs, NJ.

Wertz, J. R. (ed.). 1984. *Spacecraft Attitude Determination and Control.* D. Reidel, Holland.

Chapter

4

Radio Wave Propagation

4.1 Introduction

A signal traveling between an earth station and a satellite must pass through the earth's atmosphere, including the ionosphere, as shown in Fig. 4.1, and this can introduce certain impairments, which are summarized in Table 4.1. Some of the more important of these impairments will be described in this chapter.

4.2 Atmospheric Losses

Losses occur in the earth's atmosphere as a result of energy absorption by the atmospheric gases. These losses are treated quite separately from those which result from adverse weather conditions, which of course are also atmospheric losses. To distinguish between these, the weather-related losses are referred to as *atmospheric attenuation* and the absorption losses simply as *atmospheric absorption*.

The atmospheric absorption loss varies with frequency, as shown in Fig. 4.2. The figure is based on statistical data (CCIR Report 719-1, 1982). Two absorption peaks will be observed, the first one at a frequency of 22.3 GHz, resulting from resonance absorption in water vapor (H_2O), and the second one at 60 GHz, resulting from resonance absorption in oxygen (O_2). However, at frequencies well clear of these peaks, the absorption is quite low. The graph in Fig. 4.2 is for vertical incidence, that is, for an elevation angle of 90° at the earth-station antenna. Denoting this value of absorption loss as $[\mathrm{AA}]_{90}$ decibels, then for elevation angles down to 10°, an approximate formula for the absorption loss in decibels is (CCIR Report 719-1, 1982)

$$[\mathrm{AA}] = [\mathrm{AA}]_{90} \operatorname{cosec} El \tag{4.1}$$

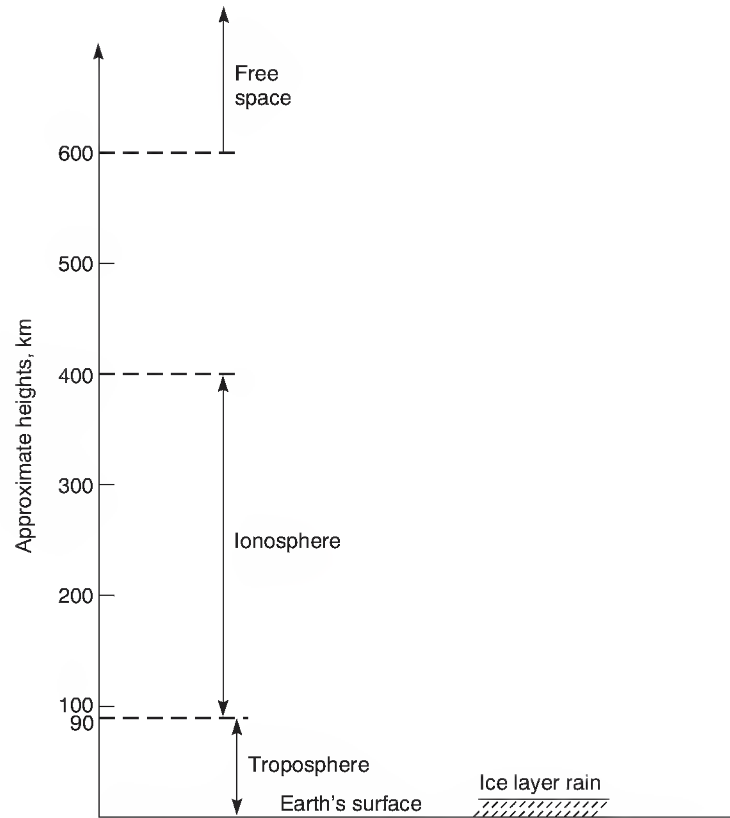


Figure 4.1 Layers in the earth's atmosphere.

where *El* is the angle of elevation. An effect known as *atmospheric scintillation* can also occur. This is a fading phenomenon, the fading period being several tens of seconds (Miya, 1981). It is caused by differences in the atmospheric refractive index, which in turn results in focusing and defocusing of the radio waves, which follow different ray paths through the atmosphere. It may be necessary to make an allowance for atmospheric scintillation, through the introduction of a fade margin in the link power-budget calculations.

4.3 Ionospheric Effects

Radio waves traveling between satellites and earth stations must pass through the ionosphere. The ionosphere is the upper region of the earth's atmosphere, which has been ionized, mainly by solar radiation. The

TABLE 4.1 Propagation Concerns for Satellite Communications Systems

Propagation impairment	Physical cause	Prime importance
Attenuation and sky noise increases	Atmospheric gases, cloud, rain	Frequencies above about 10 GHz
Signal depolarization	Rain, ice crystals	Dual-polarization systems at C and Ku bands (depends on system configuration)
Refraction, atmospheric multipath	Atmospheric gases	Communication and tracking at low elevation angles
Signal scintillations	Tropospheric and ionospheric refractivity fluctuations	Tropospheric at frequencies above 10 GHz and low-elevation angles; ionospheric at frequencies below 10 GHz
Reflection multipath, blockage	Earth's surface, objects on surface	Mobile satellite services
Propagation delays, variations	Troposphere, ionosphere	Precise timing and location systems; *time division multiple access* (TDMA) systems
Intersystem interference	Ducting, scatter, diffraction	Mainly C band at present; rain scatter may be significant at higher frequencies

SOURCE: Brussard and Rogers, 1990.

free electrons in the ionosphere are not uniformly distributed but form in layers. Furthermore, clouds of electrons (known as *traveling ionospheric disturbances*) may travel through the ionosphere and give rise to fluctuations in the signal that can only be determined on a statistical basis. The effects include *scintillation, absorption, variation in the direction of arrival, propagation delay, dispersion, frequency change,* and *polarization rotation* (CCIR Report 263-5, 1982). All these effects decrease as frequency increases, most in inverse proportion to the frequency squared, and only the polarization rotation and scintillation effects are of major concern for satellite communications. Polarization rotation is described in Sec. 5.5.

Ionospheric scintillations are variations in the amplitude, phase, polarization, or angle of arrival of radio waves. They are caused by irregularities in the ionosphere which change with time. The main effect of scintillations is fading of the signal. The fades can be quite severe, and they may last up to several minutes. As with fading caused by atmospheric scintillations, it may be necessary to include a fade margin in the link power-budget calculations to allow for ionospheric scintillation.

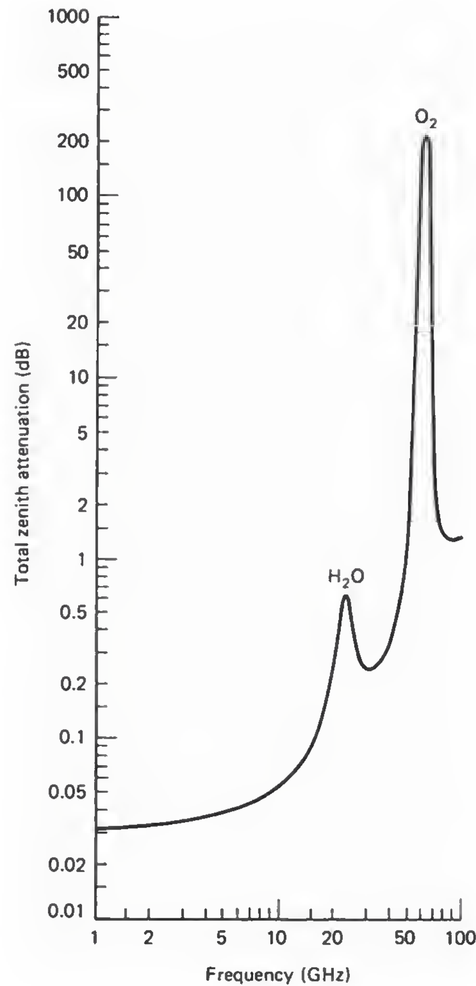


Figure 4.2 Total zenith attenuation at ground level: pressure = 1 atm, temperature = 20°C, and water vapor = 7.5 g/m^3. (*Adapted from CCIR Report 719-2, with permission from International Telecommunication Union.*)

4.4 Rain Attenuation

Rain attenuation is a function of *rain rate*. By rain rate is meant the rate at which rainwater would accumulate in a rain gauge situated at the ground in the region of interest (e.g., at an earth station). In calculations relating to radio wave attenuation, the rain rate is measured in millimeters per hour. Of interest is the percentage of time that specified values are exceeded. The time percentage is usually that of a year; for example, a rain rate of 0.001 percent means that the rain rate would be exceeded for 0.001 percent of a year, or about 5.3 min during any one year. In this case the rain rate would be denoted by $R_{0.001}$. In general,

the percentage time is denoted by p and the rain rate by R_p. The *specific attenuation* α is

$$\alpha = aR_p^b \text{ dB/km} \tag{4.2}$$

where a and b depend on frequency and polarization. Values for a and b are available in tabular form in a number of publications. The values in Table 4.2 have been abstracted from Table 4.3 of Ippolito (1986). The subscripts h and v refer to horizontal and vertical polarizations respectively.

Once the specific attenuation is found, the total attenuation is determined as

$$A = \alpha L \text{ dB} \tag{4.3}$$

where L is the *effective path length* of the signal through the rain. Because the rain density is unlikely to be uniform over the actual path length, an effective path length must be used rather than the actual (geometric) length. Figure 4.3 shows the geometry of the situation. The geometric, or slant, path length is shown as L_S. This depends on the antenna angle of elevation θ and the *rain height* h_R, which is the height at which freezing occurs. Figure 4.4 shows curves for h_R for different climatic zones. In this figure, three methods are labeled: Method 1—*maritime climates*; Method 2—*tropical climates*; Method 3—*continental climates*. For the last, curves are shown for p values of 0.001, 0.01, 0.1, and 1 percent.

For small angles of elevation ($El < 10°$), the determination of L_S is complicated by earth curvature (see CCIR Report 564-2, 1982). However,

TABLE 4.2 Specific Attenuation Coefficients

Frequency, GHz	a_h	a_v	b_h	b_v
1	0.0000387	0.0000352	0.912	0.88
2	0.000154	0.000138	0.963	0.923
4	0.00065	0.000591	1.121	1.075
6	0.00175	0.00155	1.308	1.265
7	0.00301	0.00265	1.332	1.312
8	0.00454	0.00395	1.327	1.31
10	0.0101	0.00887	1.276	1.264
12	0.0188	0.0168	1.217	1.2
15	0.0367	0.0335	1.154	1.128
20	0.0751	0.0691	1.099	1.065
25	0.124	0.113	1.061	1.03
30	0.187	0.167	1.021	1

SOURCE: Ippolito, 1986, p. 46.

TABLE 4.3 Reduction Factors

For $p = 0.001\%$	$r_{0.001} = \dfrac{10}{10 + L_G}$
For $p = 0.01\%$	$r_{0.01} = \dfrac{90}{90 + 4L_G}$
For $p = 0.1\%$	$r_{0.1} = \dfrac{180}{180 + L_G}$
For $p = 1\%$	$r_1 = 1$

SOURCE: Ippolito, 1986.

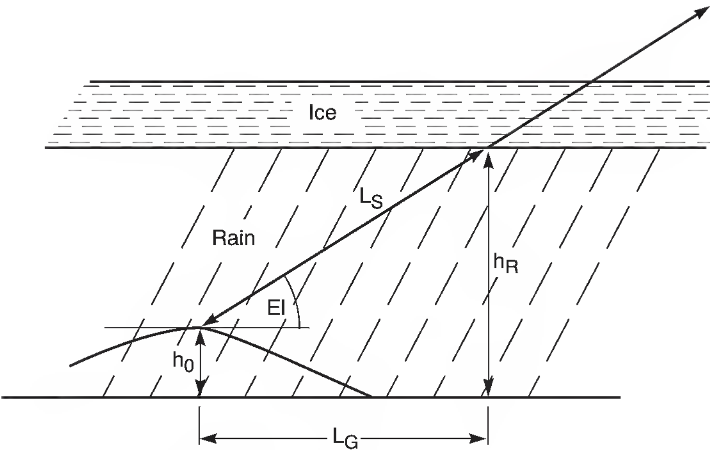


Figure 4.3 Path length through rain.

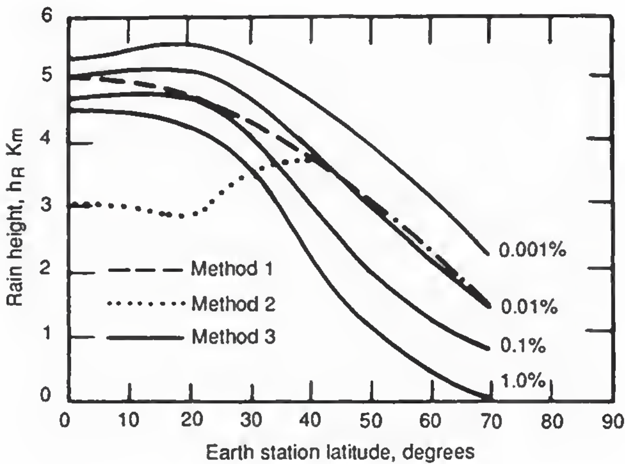


Figure 4.4 Rain height as a function of earth-station latitude for different climatic zones.

for $El \geqslant 10°$ a flat earth approximation may be used, and from Fig. 4.3 it is seen that

$$L_S = \frac{h_R - h_0}{\sin El} \tag{4.4}$$

The effective path length is given in terms of the slant length by

$$L = L_S r_p \tag{4.5}$$

where r_p is a *reduction factor* which is a function of the percentage time p and L_G, the horizontal projection of L_S. From Fig. 4.3 the horizontal projection is seen to be

$$L_G = L_S \cos El \tag{4.6}$$

The reduction factors are given in Table 4.3.

With all these factors together into one equation, the rain attenuation in decibels is given by

$$A_p = aR_p^b L_S r_p \text{ dB} \tag{4.7}$$

Interpolation formulas which depend on the climatic zone being considered are available for values of p other than those quoted earlier (see, e.g., Ippolito, 1986). Polarization shifts resulting from rain are described in Sec. 5.6.

Example 4.1 Calculate, for a frequency of 12 GHz and for horizontal and vertical polarizations, the rain attenuation which is exceeded for 0.01 percent of the time in any year, for a point rain rate of 10 mm/h. The earth station altitude is 600 m, and the antenna elevation angle is 50°. The rain height is 3 km.

Solution The given data follows. Because the CCIR formula contains hidden conversion factors, units will not be attached to the data, and it is understood that all lengths and heights are in kilometers, and rain rate is in millimeters per hour.

$$El = 50°;\ h_0 = 0.6;\ h_r = 3;\ R_{01} = 10$$

From Eq. (4.4):

$$L_S = \frac{h_R - h_0}{\sin El}$$
$$= \frac{3 - 0.6}{\sin 50°}$$
$$= 3.133 \text{ km}$$

From Eq. (4.6):

$$L_G = L_S \cos El$$
$$= 3.133 \cos 50°$$
$$= 2.014 \text{ km}$$

From Table 4.3, the reduction factor is

$$r_{01} = \frac{90}{90 + 4L_G}$$
$$= 0.9178$$

For horizontal polarization, from Table 3.2 at $f = 12$ GHz; $a_h = 0.0188$; $b_h = 1.217$
From Eq. (4.7):

$$A_p = a_h R_{01}^{b_h} L_S r_{01}$$
$$= 0.0188 \times 10^{1.217} \times 3.133 \times 0.9178$$
$$= \underline{\underline{0.891 \text{ dB}}}$$

For vertical polarization, from Table 3.2 at $f = 12$ GHz; $a_v = 0.0168$; $b_v = 1.2$

$$A_p = a_v R_{01}^{b_h} L_S r_{01}$$
$$= 0.0168 \times 10^{1.2} \times 3.133 \times 0.9178$$
$$= \underline{\underline{0.766 \text{ dB}}}$$

The corresponding equations for circular polarization are

$$a_c = \frac{a_h + a_v}{2} \tag{4.8a}$$

$$b_c = \frac{a_h b_h + a_v b_v}{2a_c} \tag{4.8b}$$

The attenuation for circular polarization is compared with that for linear polarization in the following example.

Example 4.2 Repeat Example 4.1 for circular polarization.

Solution From Eq. (4.8*a*):

$$a_c = \frac{a_h + a_v}{2}$$
$$= \frac{0.0188 + 0.0168}{2}$$
$$= 0.0178$$

From Eq. (4.8*b*):

$$b_c = \frac{a_h b_h + a_v b_v}{2a_c}$$

$$= \frac{0.0188 \times 1.217 + 0.0168 \times 1.2}{2 \times 0.0178}$$

$$= 1.209$$

From Eq. (4.7):

$$A_p = a_c R_{01}^{b_c} L_S r_{01}$$

$$= 0.0178 \times 10^{1.209} \times 3.133 \times 0.9178$$

$$= \underline{\underline{0.828 \text{ dB}}}$$

4.5 Other Propagation Impairments

Hail, ice, and snow have little effect on attenuation because of the low water content. Ice can cause depolarization, described briefly in Chap. 5. The attenuation resulting from clouds can be calculated as that for rain (Ippolito, 1986, p. 56), although the attenuation is generally much less. For example, at a frequency of 10 GHz and a water content of 0.25 g/m^3, the specific attenuation is about 0.05 dB/km and about 0.2 dB/km for a water content of 2.5 g/m^3.

4.6 Problems and Exercises

4.1. With reference to Table 4.1, identify the propagation impairments which most affect transmission in the C band.

4.2. Repeat Prob. 4.1 for Ku-band transmissions.

4.3. Calculate the approximate value of atmospheric attenuation for a satellite transmission at 14 GHz, for which the angle of elevation of the earth-station antenna is 15°.

4.4. Calculate the approximate value of atmospheric attenuation for a satellite transmission at 6 GHz, for which the angle of elevation of the earth-station antenna is 30°.

4.5. Describe the major effects the ionosphere has on the transmission of satellite signals at frequencies of (*a*) 4 GHz and (*b*) 12 GHz.

4.6. Explain what is meant by *rain rate* and how this is related to specific attenuation.

4.7. Compare the specific attenuations for vertical and horizontal polarization at a frequency of 4 GHz and a point rain rate of 8 mm/h which is exceeded for 0.01 percent of the year.

4.8. Repeat Prob. 4.7 for a frequency of 12 GHz.

4.9. Explain what is meant by *effective path length* in connection with rain attenuation.

4.10. For a satellite transmission path, the angle of elevation of the earth station antenna is 35°, and the earth station is situated at mean sea level. The signal is vertically polarized at a frequency of 18 GHz. The rain height is 1 km, and a rain rate of 10 mm/h is exceeded for 0.001 percent of the year. Calculate the rain attenuation under these conditions.

4.11. Repeat Prob. 4.10 when the rain rate of 10 mm/h is exceeded (*a*) 0.01 percent and (*b*) 0.1 percent of the year.

4.12. Given that for a satellite transmission $El = 22°$, $R_{0.01} = 15$ mm/h, $h_0 = 600$ m, $h_R = 1500$ m, and horizontal polarization is used, calculate the rain attenuation for a signal frequency of 14 GHz.

4.13. Determine the specific attenuation for a circularly polarized satellite signal at a frequency of 4 GHz, where a point rain rate of 8 mm/h is exceeded for 0.01 percent of the year.

4.14. A circularly polarized wave at a frequency of 12 GHz is transmitted from a satellite. The point rain rate for the region is $R_{0.01} = 13$ mm/h. Calculate the specific attenuation.

4.15. Given that for Prob. 4.13 the earth station is situated at altitude 500 m and the rain height is 2 km, calculate the rain attenuation. The angle of elevation of the path is 35°.

4.16. Given that for Prob. 4.14 the earth station is situated at altitude 200 m and the rain height is 2 km, calculate the rain attenuation. The angle of elevation of the path is 25°.

References

Brussard, G., and D. V. Rogers. 1990. "Propagation Considerations in Satellite Communication Systems." *Proc. IEEE,* Vol. 78, No. 7, July, pp. 1275–1282.

CCIR Report 263-5. 1982. "Ionospheric Effects upon Earth-Space Propagation." *15th Plenary Assembly,* Vol. VI, Geneva, pp. 124–146.

CCIR Report 564-2. 1982. "Propagation Data Required for Space Telecommunication System." *15 Plenary Assembly,* Vol. IX, Part 1, Geneva.

CCIR Report 719-1. 1982. "Attenuation by Atmospheric Gases." *15th Plenary Assembly*, Vol. V, Geneva, pp. 138–150.

Ippolito, L. J. 1986. *Radiowave Propagation in Satellite Communications*. Van Nostrand Reinhold, New York.

Miya, K. (ed.). 1981. *Satellite Communications Technology*. KDD Engineering and Consulting, Japan.

Chapter

5

Polarization

5.1 Introduction

In the *far field zone* of a transmitting antenna, the radiated wave takes on the characteristics of a *transverse electromagnetic* (TEM) wave. Far field zone refers to distances greater than $2D^2/\lambda$ from the antenna, where D is the largest linear dimension of the antenna and λ is the wavelength. For a parabolic antenna of 3 m diameter transmitting a 6-GHz wave ($\lambda = 5$ cm), the far field zone begins at approximately 360 m. The TEM designation is illustrated in Fig. 5.1, where it can be seen that both the magnetic field **H** and the electric field **E** are transverse to the direction of propagation, denoted by the propagation vector **k.**

E, H, and **k** represent vector quantities, and it is important to note their relative directions. When one looks along the direction of propagation, the rotation from **E** to **H** is in the direction of rotation of a right-hand-threaded screw, and the vectors are said to form a *right-hand set.* The wave always retains the directional properties of the right-hand set, even when reflected, for example. One way of remembering how the right-hand set appears is to note that the letter E comes before H in the alphabet and rotation is from **E** to **H** when looking along the direction of propagation.

At great distances from the transmitting antenna, such as are normally encountered in radio systems, the TEM wave can be considered to be plane. This means that the **E** and **H** vectors lie in a plane, which is at right angles to the vector **k.** The vector **k** is said to be normal to the plane. The magnitudes are related by $E = HZ_0$, where $Z_0 = 120\pi\Omega$.

The direction of the line traced out by the tip of the electric field vector determines the *polarization* of the wave. Keep in mind that the electric and magnetic fields are varying as functions of time. The magnetic field varies exactly in phase with the electric field, and its amplitude is proportional to the electric field amplitude, so it is only necessary

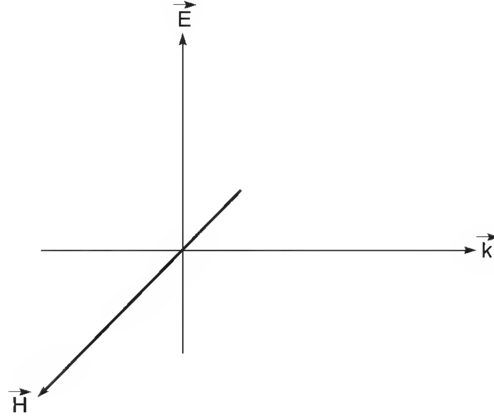


Figure 5.1 Vector diagram for a transverse electromagnetic (TEM) wave.

to consider the electric field in this discussion. The tip of the **E** vector may trace out a straight line, in which case the polarization is referred to as *linear.* Other forms of polarization, specifically elliptical and circular, will be introduced later.

In the early days of radio, there was little chance of ambiguity in specifying the direction of polarization in relation to the surface of the earth. Most transmissions utilized linear polarization and were along terrestrial paths. Thus *vertical polarization* meant that the electric field was perpendicular to the earth's surface, and *horizontal polarization* meant that it was parallel to the earth's surface. Although the terms vertical and horizontal are used with satellite transmissions, the situation is not quite so clear. A linear polarized wave transmitted by a geostationary satellite may be designated vertical if its electric field is parallel to the earth's polar axis, but even so the electric field will be parallel to the earth at the equator. This situation will be clarified shortly.

Suppose for the moment that horizontal and vertical are taken as the x and y axes of a right-hand set, as shown in Fig. 5.2*a*. A vertically polarized electric field can be described as

$$\mathbf{E}_y = \hat{a}_y E_y \sin \omega t \tag{5.1}$$

where $\hat{a}_y$ is the unit vector in the vertical direction and E_y is the *peak value* or *amplitude* of the electric field. Likewise, a horizontally polarized wave could be described by

$$\mathbf{E}_x = \hat{a}_x E_x \sin \omega t \tag{5.2}$$

These two fields would trace out the straight lines shown in Fig. 5.2*b*. Now consider the situation where both fields are present simultaneously.

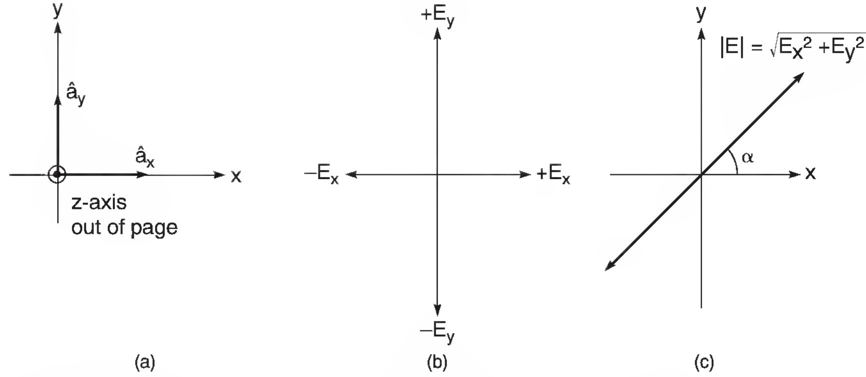


Figure 5.2 Horizontal and vertical components of linear polarization.

These would add vectorially, and the resultant would be a vector **E** (Fig. 5.2*c*) of amplitude $\sqrt{E_x^2 + E_y^2}$, at an angle to the horizontal given by

$$\alpha = \arctan \frac{E_y}{E_x} \tag{5.3}$$

E varies sinusoidally in time in the same manner as the individual components. It is still linearly polarized but cannot be classified as simply horizontal or vertical. Arguing back from this, it is evident that **E** can be resolved into vertical and horizontal components, a fact which is of great importance in practical transmission systems. The power in the resultant wave is proportional to the voltage $\sqrt{E_x^2 + E_y^2}$, squared, which is $E_x^2 + E_y^2$. In other words, the power in the resultant wave is the sum of the powers in the individual waves, which is to be expected.

More formally, $\mathbf{E}_y$ and $\mathbf{E}_x$ are said to be *orthogonal*. The dictionary definition of orthogonal is at *right angles*, but a wider meaning will be attached to the word later.

Consider now the situation where the two fields are equal in amplitude (denoted by E), but one leads the other by 90° in phase. The equations describing these are

$$\mathbf{E}_y = \hat{a}_y E \sin \omega t \tag{5.4a}$$

$$\mathbf{E}_x = \hat{a}_x E \cos \omega t \tag{5.4b}$$

Applying Eq. (5.3) in this case yields $\alpha = \omega t$. The tip of the resultant electric field vector traces out a circle, as shown in Fig. 5.3*a*, and the resultant wave is said to be *circularly polarized*. The amplitude of the resultant vector is E. The resultant field in this case does not go through zero. At $\omega t = 0$, the y component is zero and the x component is E. At

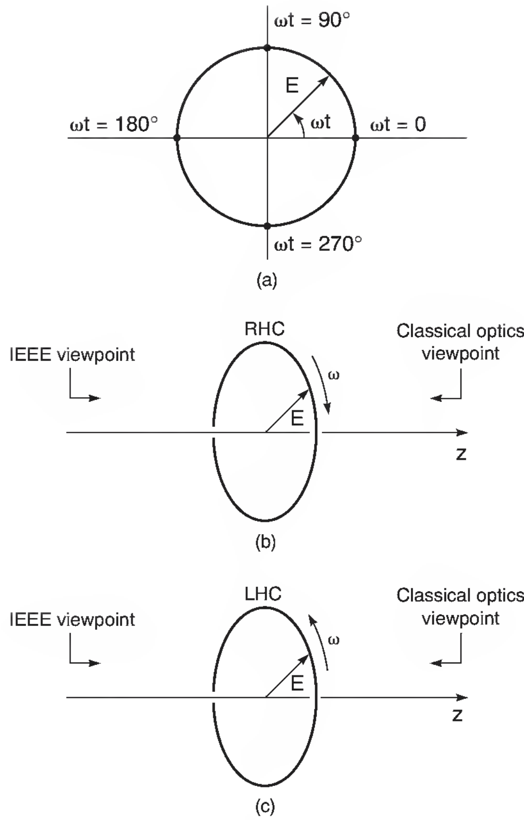


Figure 5.3 Circular polarization.

$\omega t = 90°$, the y component is E and the x component is zero. Compare this with the linear polarized case where at $\omega t = 0$, both the x and y components are zero, and at $\omega t = 90°$, both components are maximum at E. Because the resultant does not vary in time, the power must be found by adding the powers in the two linear polarized, sinusoidal waves. This gives a resultant proportional to $2E^2$.

The direction of circular polarization is defined by the sense of rotation of the electric vector, but this also requires that the way the vector is viewed must be specified. The *Institute of Electrical and Electronics Engineers* (IEEE) defines *right-hand circular* (RHC) *polarization* as a rotation in the clockwise direction when the wave is viewed along the direction of propagation, that is, when viewed from "behind," as shown in Fig. 5.3*b*. *Left-hand circular* (LHC) *polarization* is when the rotation is in the counterclockwise direction when viewed along the direction of propagation, as shown in Fig. 5.3*c*. LHC and RHC polarizations are orthogonal. The direction of propagation is along the $+z$ axis.

As a caution it should be noted that the classical optics definition of circular polarization is just the opposite of the IEEE definition. The IEEE definition will be used throughout this text.

For a right-hand set of axes (Fig. 5.1) and with propagation along the $+z$ axis, then when viewed along the direction of propagation (from "behind") and with the $+y$ axis directed upward, the $+x$ axis will be directed toward the left. Consider now Eq. (5.4). At $\omega t = 0$, E_y is 0 and E_x is a maximum at E along the $+x$ axis. At $\omega t = 90°$, E_x is zero and E_y is a maximum at E along the $+y$ axis. In other words, the resultant field of amplitude E has rotated from the $+x$ axis to the $+y$ axis, which is a clockwise rotation when viewed along the direction of propagation. Equation (5.4) therefore represents RHC polarization.

Given that Eq. (5.4) represents RHC polarization, it is left as an exercise to show that the following equations represent LHC polarization:

$$\mathbf{E}_y = \hat{a}_y E \sin \omega t \tag{5.5a}$$

$$\mathbf{E}_x = -\hat{a}_x E \cos \omega t \tag{5.5b}$$

In the more general case, a wave may be *elliptically polarized*. This occurs when the two linear components are

$$\mathbf{E}_y = \hat{a}_y E_y \sin \omega t \tag{5.6a}$$

$$\mathbf{E}_x = \hat{a}_x E_x \sin(\omega t + \delta) \tag{5.6b}$$

Here, E_y and E_x are not equal in general, and δ is a fixed phase angle. It is left as an exercise for the student to show that when $E_y = 1$, $E_x = 1/3$, and $\delta = 30°$, the polarization ellipse is as shown in Fig. 5.4.

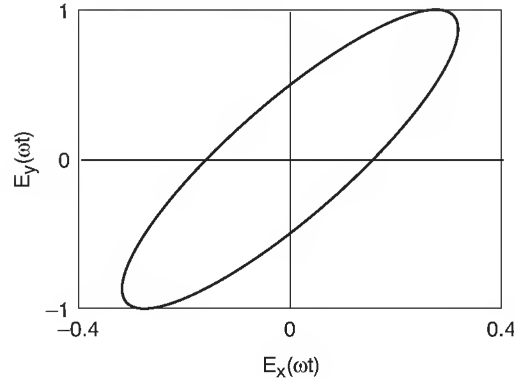


Figure 5.4 Elliptical polarization.

The *axial ratio* of an elliptical polarized wave is the ratio of major axis to minor axis of the ellipse. Orthogonal elliptical polarization occurs when a wave has the same value of axial ratio but opposite sense of rotation.

Satellite communications links use linear polarization and circular polarization, but transmission impairments can change the polarization to elliptical in each case. Some of these impairments, relating to the transmission medium, are described in Secs. 5.5, 5.6, and 5.7, and the influence of the antenna structure on polarization is described in Chap. 6. Antennas are covered in detail in Chap. 6, but at this stage the relationship of the antenna to the polarization type will be defined.

5.2 Antenna Polarization

The polarization of a transmitting antenna is defined by the polarization of the wave it transmits. Thus a horizontal dipole would produce a horizontally polarized wave. Two dipoles mounted close together symmetrically and at right angles to each other would produce a circularly polarized wave if fed with currents equal in amplitude but differing in phase by 90°. This is shown by Eqs. (5.4) and (5.5). Note that because of the symmetry of the circular polarization, the dipoles need not lie along the horizontal and vertical axes; they just need to be spatially at right angles to each other. The terms *horizontal* and *vertical* are used for convenience.

The polarization of a receiving antenna has to be aligned to that of the wave for maximum power transfer. Taking again the simple dipole as an example, a vertical dipole will receive maximum signal from a vertically polarized wave. Figure 5.5 illustrates this. In Fig. 5.5*a* the dipole

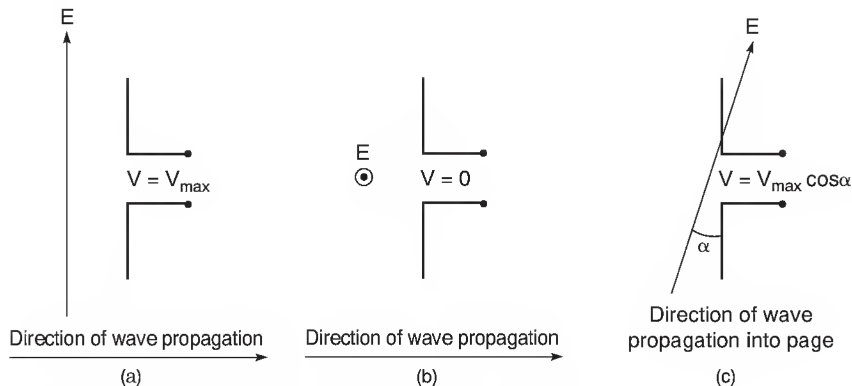


Figure 5.5 Linear polarization relative to a receiving dipole.

is parallel to the electric field E, and hence the induced voltage V will be a maximum, denoted by V_{max}. In Fig. 5.5*b* the dipole is at right angles to the electric field, and the induced voltage is zero. In Fig. 5.5*c* the dipole lies in the plane of polarization (the wavefront) but is at some angle α to the electric field. The induced voltage will be given by

$$V = V_{max} \cos \alpha \tag{5.7}$$

Note that for Eq. (5.7) to apply, the dipole has to lie in the same plane as E (the wavefront). If the dipole is inclined at some angle θ to the wavefront, the received signal is reduced further by the radiation pattern of the antenna. This is described more fully in Sec. 6.6. The reciprocity theorem for antennas (see Sec. 6.2) ensures that an antenna designed to transmit in a given polarization will receive maximum power from a wave with that polarization. An antenna designed for a given sense of polarization will receive no energy from a wave with the orthogonal polarization. Figures 5.5*a* and *b* illustrate the specific case where the desired signal is vertically polarized and the orthogonal signal is horizontally polarized. However, as mentioned above, certain impairments can result in a loss of polarization discrimination, discussed in later sections.

The combined power received by the two crossed dipoles will be maximum when the incoming wave is circularly polarized. The average power received from a sinusoidal wave is proportional to the square of the amplitude. Thus, for a circularly polarized wave given by either of Eq. (5.4) or Eq. (5.5), the power received from each component is proportional to E^2, and the total power is twice that of one component alone. The crossed dipoles would receive this total. A single dipole will always receive a signal from a circularly polarized wave, but at a loss of 3 dB. This is so because the single dipole will respond only to one of the linear components, and hence the received power will be half that of the crossed dipoles. Again, because of the symmetry of the circularly polarized wave, the dipole need only lie in the plane of polarization; its orientation with respect to the *xy*-axes is not a factor.

A grid of parallel wires will reflect a linear polarized wave when the electric field is parallel to the wires, and it will transmit the orthogonal wave. This is illustrated in Fig. 5.6. This is used in one type of *dual polarized antenna*, illustrated in Fig. 5.7. Here, the grid allows the wave, the electric field of which is transverse to the wires to pass through, whereas it reflects the parallel (E_v) wave. The reflector behind the grid reflects the wave that passes through. Thus two orthogonal, linearly polarized waves, having high polarization isolation (see Sec. 5.4) are transmitted from the antenna system. Some details of the construction of this type of antenna will be found in Maral and Bousquet (1998).

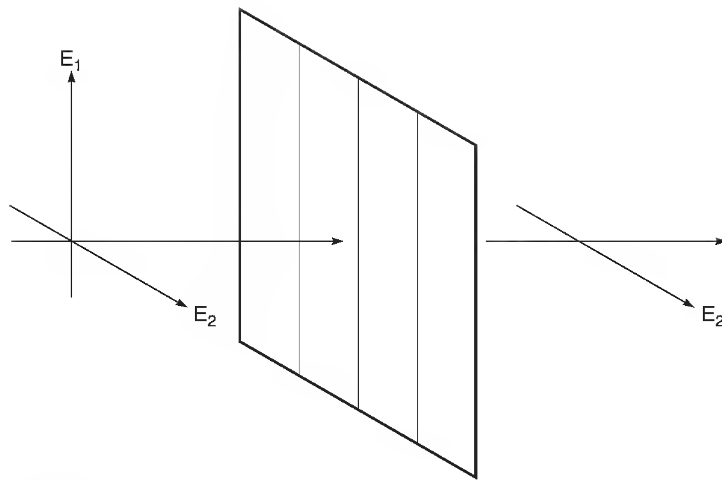


Figure 5.6 A wire grid polarizer.

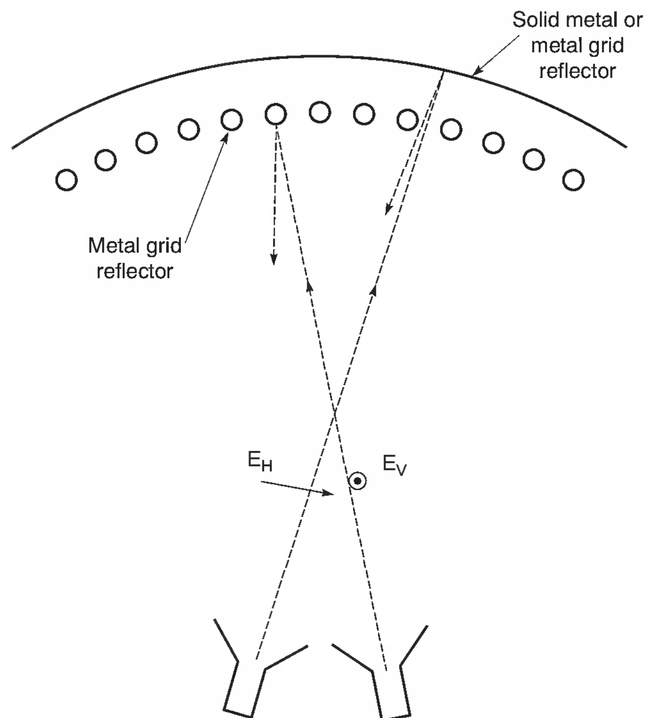


Figure 5.7 A wire grid polarizer used in a dual-polarized antenna.

5.3 Polarization of Satellite Signals

As mentioned above, the directions "horizontal" and "vertical" are easily visualized with reference to the earth. Consider, however, the situation where a geostationary satellite is transmitting a linear polarized wave. In this situation, the usual definition of horizontal polarization is where the electric field vector is parallel to the equatorial plane, and vertical polarization is where the electric field vector is parallel to the earth's polar axis. It will be seen that at the subsatellite point on the equator, both polarizations will result in electric fields that are parallel to the local horizontal plane, and care must be taken therefore not to use "horizontal" as defined for terrestrial systems. For other points on the earth's surface within the footprint of the satellite beam, the polarization vector (the unit vector in the direction of the electric field) will be at some angle relative to a reference plane. Following the work of Hogg and Chu (1975), the reference plane will be taken to be that which contains the direction of propagation and the local gravity direction (a "plumb line"). This is shown in Fig. 5.8.

With the propagation direction denoted by **k** and the local gravity direction at the ground station by **r,** the direction of the normal to the reference plane is given by the vector cross-product:

$$\mathbf{f} = \mathbf{k} \times \mathbf{r} \tag{5.8}$$

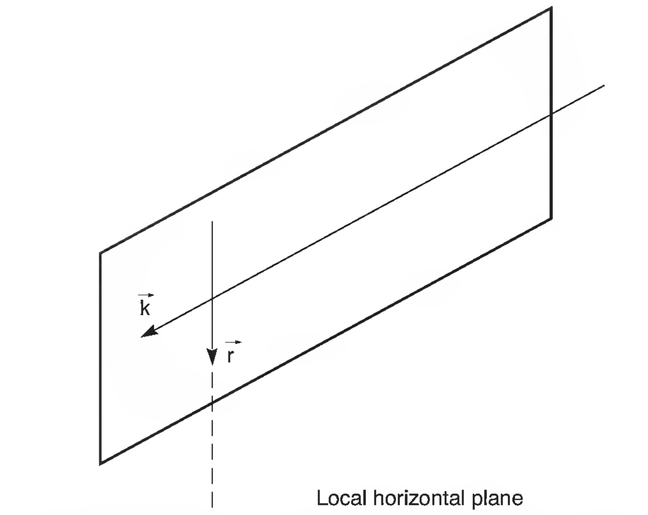


Figure 5.8 The reference plane for the direction of propagation and the local gravity direction.

With the unit polarization vector at the earth station denoted by **p**, the angle between it and **f** is obtained from the vector dot product as

$$\eta = \arccos \frac{\mathbf{p} \cdot \mathbf{f}}{|\mathbf{f}|} \tag{5.9}$$

Since the angle between a normal and its plane is 90°, the angle between **p** and the reference plane is $\xi = |90^\circ - \eta|$ and

$$\xi = \arcsin \frac{\mathbf{p} \cdot \mathbf{f}}{|\mathbf{f}|} \tag{5.10}$$

This is the desired angle. Keep in mind that the polarization vector is always at right angles to the direction of propagation.

The next step is to relate the polarization vector **p** to the defined polarization at the satellite. Let unit vector **e** represent the defined polarization at the satellite. For vertical polarization, **e** lies parallel to the earth's N-S axis. For horizontal polarization, **e** lies in the equatorial plane at right angles to the geostationary radius a_{GSO} to the satellite. A cross-product vector can be formed,

$$\mathbf{g} = \mathbf{k} \times \mathbf{e} \tag{5.11}$$

where **g** is normal to the plane containing **e** and **k**, as shown in Fig. 5.9. The cross-product of **g** with **k** gives the direction of the polarization in this plane. Denoting this cross-product by **h** gives

$$\mathbf{h} = \mathbf{g} \times \mathbf{k} \tag{5.12}$$

The unit polarization vector at the earth station is therefore given by

$$\mathbf{p} = \frac{\mathbf{h}}{|\mathbf{h}|} \tag{5.13}$$

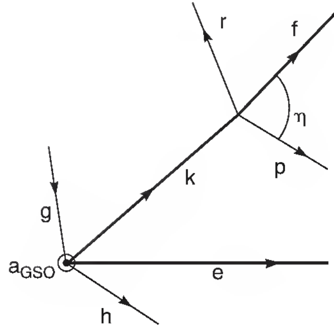


Figure 5.9 Vectors $\mathbf{g} = \mathbf{k} \times \mathbf{e}$ and $\mathbf{h} = \mathbf{g} \times \mathbf{h}$.

All these vectors can be related to the known coordinates of the earth station and satellite shown in Fig. 5.10. With the longitude of the satellite as the reference, the satellite is positioned along the positive x axis at

$$x_s = a_{\mathrm{GSO}} \tag{5.14}$$

The coordinates for the earth-station position vector **R** are (ignoring the slight difference between geodetic and geocentric latitudes and

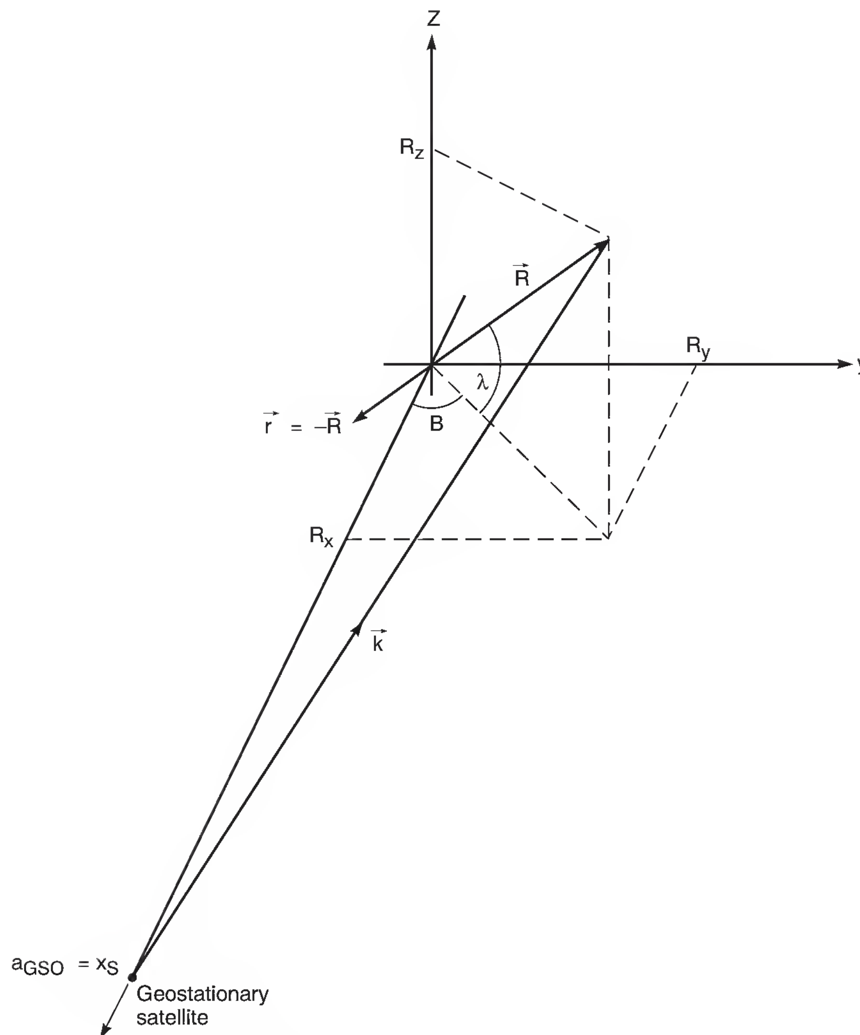


Figure 5.10 Vectors **k** and **R** in relation to satellite and earth station positions.

assuming the earth station to be at mean sea level)

$$R_x = R\cos\lambda \cos B \tag{5.15a}$$

$$R_y = R\cos\lambda \sin B \tag{5.15b}$$

$$R_z = R\sin\lambda \tag{5.15c}$$

where $B = \phi_E - \phi_{SS}$ as defined in Eq. (3.8).

The local gravity direction is $\mathbf{r} = -\mathbf{R}$. The coordinates for the direction of propagation $\mathbf{k}$ are

$$k_x = R_x - a_{GSO} \tag{5.16a}$$

$$k_y = R_y \tag{5.16b}$$

$$k_z = R_z \tag{5.16c}$$

Calculation of the polarization angle is illustrated in the following example.

Example 5.1 A geostationary satellite is stationed at 105°W and transmits a vertically polarized wave. Determine the angle of polarization at an earth station at latitude 18°N longitude 73°W.

Solution Given data:
$\lambda = 18°$; $\phi_E = -73°$; $\phi_{SS} = -105°$; $a_{GSO} = 42164$ km; R = 6371 km (spherical earth of mean radius R assumed).
Eq. (3.8) gives:

$$B = \phi_E - \phi_{SS} = 32°$$

Applying Eq. (5.15), the geocentric-equatorial coordinates for the earth station position vector are:

$$R_x = R\cos\lambda \cos B$$
$$= 6371 \cos 18° \cos 32°$$
$$= 5138.48 \text{ km}$$

$$R_y = R\cos\lambda \sin B$$
$$= 6371 \cos 18° \sin 32°$$
$$= 3210.88 \text{ km}$$

$$R_z = R\sin\lambda$$
$$= 6371 \sin 18°$$
$$= 1968.75 \text{ km}$$

The coordinates for the local gravity direction, obtained from $\mathbf{r} = -\mathbf{R}$ are

$$\mathbf{r} = -\begin{bmatrix} 5138.48 \\ 3210.88 \\ 1968.75 \end{bmatrix} \text{km}$$

From Eq. (5.16), the geocentric-equatorial coordinates for the propagation direction are

$$\mathbf{k} = \begin{bmatrix} R_x - a_{\text{GSO}} \\ R_y \\ R_z \end{bmatrix} = \begin{bmatrix} -37025.5 \\ 3210.88 \\ 1968.75 \end{bmatrix} \text{km}$$

For vertical polarization at the satellite, the geocentric-equatorial coordinates for the polarization vector are $x = 0$, $y = 0$, and $z = 1$:

$$\mathbf{e} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

The vector cross products can be written in determinant form, where $\mathbf{a}_x$, $\mathbf{a}_y$, $\mathbf{a}_z$, are the unit vectors along the x, y, z, axes. Thus, Eq. (5.8) is

$$\begin{aligned} \mathbf{f} &= \mathbf{k} \times \mathbf{r} \\ &= -\begin{bmatrix} \mathbf{a}_x & \mathbf{a}_y & \mathbf{a}_z \\ -37025.5 & 3210.88 & 1968.75 \\ 5138.48 & 3210.88 & 1968.75 \end{bmatrix} \\ &= \mathbf{a}_x 0 - \mathbf{a}_y 8.3 \cdot 10^7 + \mathbf{a}_z 1.35 \cdot 10^8 \text{ km}^2 \end{aligned}$$

From Eq. (5.11):

$$\begin{aligned} \mathbf{g} &= \mathbf{k} \times \mathbf{e} \\ &= \begin{bmatrix} \mathbf{a}_x & \mathbf{a}_y & \mathbf{a}_z \\ -37025.5 & 3210.88 & 1968.75 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \mathbf{a}_x 3210.88 + \mathbf{a}_y 37025.5 + \mathbf{a}_z 0 \text{ km} \end{aligned}$$

From Eq. (5.12):

$$\begin{aligned} \mathbf{h} &= \mathbf{g} \times \mathbf{k} \\ &= \begin{bmatrix} \mathbf{a}_x & \mathbf{a}_y & \mathbf{a}_z \\ 3210.88 & 37025.5 & 0 \\ -37025.5 & 3210.88 & 1968.75 \end{bmatrix} \\ &= \mathbf{a}_x 7.2894 \cdot 10^7 - \mathbf{a}_y 6.3214 \cdot 10^6 + \mathbf{a}_z 1.3812 \cdot 10^9 \text{ km}^2 \end{aligned}$$

The magnitude of **h** is

$$|\mathbf{h}| = \sqrt{(7.2894 \cdot 10^7)^2 + (-6.3214 \cdot 10^6)^2 + (1.3812 \cdot 10^9)^2}$$
$$= 1.383 \cdot 10^9$$

From Eq. (5.13):

$$\mathbf{p} = \frac{\mathbf{h}}{|\mathbf{h}|}$$
$$= \mathbf{a}_x 0.0527 - \mathbf{a}_y 0.0046 + \mathbf{a}_z 0.9986$$

The dot product of **p** and **f** is

$$\mathbf{p} \cdot \mathbf{f} = 0.0527 \times 0 + (-0.0046) \times 8.3 \times 10^7 + 0.9986 \times 1.35 \times 10^8$$
$$= 1.356 \times 10^8 \text{ km}^2$$

The magnitude of **f** is

$$|\mathbf{f}| = \sqrt{(-8.3 \times 10^7)^2 + (1.35 \times 10^8)^2}$$
$$= 1.588 \times 10^8 \text{ km}^2$$

and from Eq. (5.10)

$$\xi = \arcsin \frac{1.356}{1.588}$$
$$= \underline{\underline{58.64°}}$$

5.4 Cross-Polarization Discrimination

The propagation path between a satellite and earth station passes through the ionosphere, and possibly through layers of ice crystals in the upper atmosphere and rain, all of which are capable of altering the polarization of the wave being transmitted. An orthogonal component may be generated from the transmitted polarization, an effect referred to as *depolarization*. This can cause interference where orthogonal polarization is used to provide isolation between signals, as in the case of frequency reuse.

Two measures are in use to quantify the effects of polarization interference. The most widely used measure is called *cross-polarization discrimination* (XPD). Figure 5.11*a* shows how this is defined. The transmitted electric field is shown having a magnitude E_1 before it enters the medium which causes depolarization. At the receiving antenna the electric field may have two components, a *copolar* component, having magnitude E_{11}, and a *cross-polar* component, having magnitude E_{12}. The cross-polarization discrimination in decibels is defined as

$$\text{XPD} = 20 \log \frac{E_{11}}{E_{12}} \tag{5.17}$$

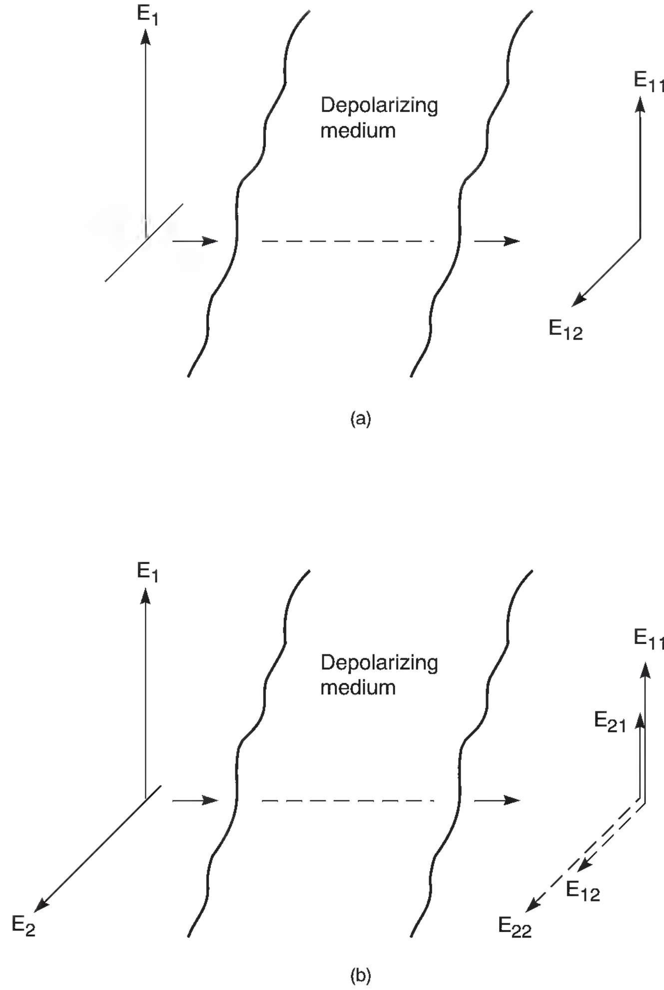


Figure 5.11 Vectors defining (a) cross-polarization discrimination (XPD), and (b) polarization isolation (I).

The second situation is shown in Fig. 5.11*b*. Here, two orthogonally polarized signals, with magnitudes E_1 and E_2, are transmitted. After traversing the depolarizing medium, copolar and cross-polar components exist for both waves. The *polarization isolation* is defined by the ratio of received copolar power to received cross-polar power and thus takes into account any additional depolarization introduced by the receiving system (Ippolito, 1986). Since received power is proportional to the square of the electric field strength, the polarization isolation in decibels is defined as

$$I = 20 \log \frac{E_{11}}{E_{21}} \tag{5.18}$$

When the transmitted signals have the same magnitudes ($E_1 = E_2$) and where the receiving system introduces negligible depolarization, then I and XPD give identical results.

For clarity, linear polarization is shown in Fig. 5.11, but the same definitions for XPD and I apply for any other system of orthogonal polarization.

5.5 Ionospheric Depolarization

The ionosphere is the upper region of the earth's atmosphere that has been ionized, mainly by solar radiation. The free electrons in the ionosphere are not uniformly distributed but form layers. Furthermore, clouds of electrons (known as *traveling ionospheric disturbances*) may travel through the ionosphere and give rise to fluctuations in the signal. One of the effects of the ionosphere is to produce a rotation of the polarization of a signal, an effect known as *Faraday rotation*.

When a linearly polarized wave traverses the ionosphere, it sets in motion the free electrons in the ionized layers. These electrons move in the earth's magnetic field, and therefore, they experience a force (similar to that which a current-carrying conductor experiences in the magnetic field of a motor). The direction of electron motion is no longer parallel to the electric field of the wave, and as the electrons react back on the wave, the net effect is to shift the polarization. The angular shift in polarization (the Faraday rotation) is dependent on the length of the path in the ionosphere, the strength of the earth's magnetic field in the ionized region, and the electron density in the region. Faraday rotation is inversely proportional to frequency squared and is not considered to be a serious problem for frequencies above about 10 GHz.

Suppose a linearly polarized wave produces an electric field E at the receiver antenna when no Faraday rotation is present. The received power is proportional to E^2. A Faraday rotation of θ_F degrees will result in the copolarized component (the desired component) of the received signal being reduced to $E_{co} = E \cos \theta_F$, the received power in this case being proportional to E_{co}^2. The *polarization loss* (PL) in decibels is

$$\begin{aligned} \mathrm{PL} &= 20 \log \frac{E_{co}}{E} \\ &= 20 \log(\cos \theta_F) \end{aligned} \tag{5.19}$$

At the same time, a cross-polar component $E_x = E \sin\theta_F$ is created, and hence the XPD is

$$\begin{aligned} \mathrm{XPD} &= 20 \log \frac{E_{co}}{E_x} \\ &= 20 \log(\cot \theta_F) \end{aligned} \tag{5.20}$$

Maximum values quoted by Miya (1981) for Faraday rotation are 9° at 4 GHz and 4° at 6 GHz. In order to counter the depolarizing effects of Faraday rotation, circular polarization may be used. With circular polarization, a Faraday shift simply adds to the overall rotation and does not affect the copolar or cross-polar components of electric field. Alternatively, if linear polarization is to be used, polarization tracking equipment may be installed at the antenna.

5.6 Rain Depolarization

The ideal shape of a raindrop is spherical, since this minimizes the energy (the surface tension) required to hold the raindrop together. The shape of small raindrops is close to spherical, but larger drops are better modeled as oblate spheroids with some flattening underneath, as a result of the air resistance. These are sketched in Fig. 5.12*a* and *b*. For vertically falling rain, the axis of symmetry of the raindrops will be parallel to the local vertical as shown in Fig. 5.12*b*, but more realistically, aerodynamic forces will cause some canting, or tilting, of the drops. Thus there will be a certain randomness in the angle of tilt as sketched in Fig. 5.12*c*.

As shown earlier, a linearly polarized wave can be resolved into two component waves, one vertically polarized and the other horizontally polarized. Consider a wave with its electric vector at some angle τ relative to the major axis of a raindrop, which for clarity is shown horizontal in Fig. 5.13. The vertical component of the electric field lies parallel to the minor axis of the raindrop and therefore encounters less water than the horizontal component. There will be a difference therefore in the attenuation and phase shift experienced by each of the electric field components. These differences are termed as the *differential attenuation and differential phase shift*, and they result in depolarization of the wave. For the situation shown in Fig. 5.13, the angle of polarization of the wave emerging from the rain is altered relative to that of the wave entering the rain. Experience has

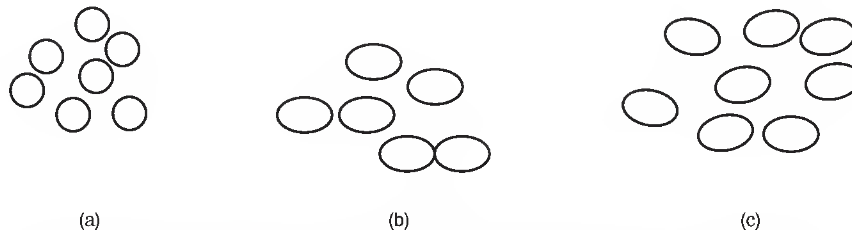


Figure 5.12 Raindrops: (*a*) small spherical, (*b*) flattening resulting from air resistance, and (*c*) angle of tilt randomized through aerodynamic force.

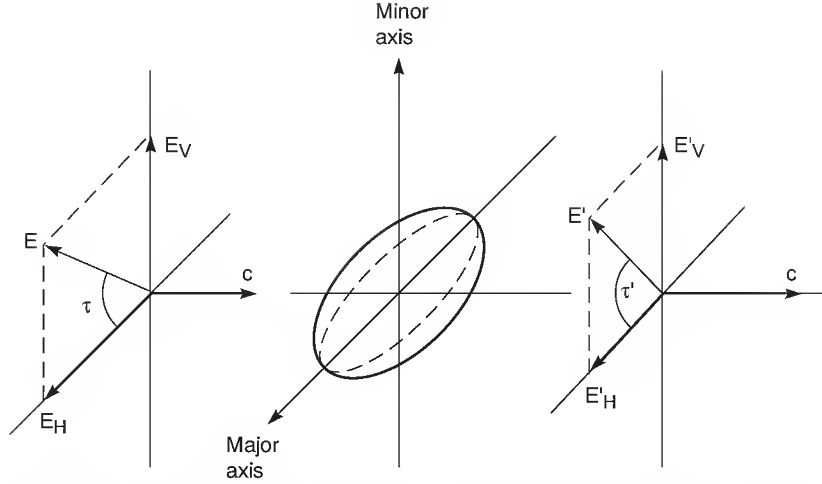


Figure 5.13 Polarization vector relative to the major and minor axes of a raindrop.

shown that the depolarization resulting from the differential phase shift is more significant than that resulting from differential attenuation.

The cross-polarization discrimination in decibels associated with rain is given to a good approximation by the empirical relationship (CCIR Report 564-2, 1982)

$$\text{XPD} = U - V \log A \tag{5.21}$$

where U and V are empirically determined coefficients and A is the rain attenuation. U, V, and A must be in decibels in this equation. The attenuation A is as determined in Sec. 4.4. The following formulas are given in the CCIR reference for U and V for the frequency range 8 to 35 GHz:

$$V = \begin{cases} 20 & \text{for } 8 \le f \le 15 \text{ GHz} \\ 23 & \text{for } 15 \le f \le 35 \text{ GHz} \end{cases} \tag{5.22a}$$

and

$$U = 30 \log f - 10 \log (0.5 - 0.4697 \cos 4\tau) - 40 \log(\cos \theta) \tag{5.22b}$$

where f is the frequency in gigahertz, θ is the angle of elevation of the propagation path at the earth station, and τ is the tilt angle of the polarization relative to the horizontal. For circular polarization $\tau = 45°$. As shown earlier, for a satellite transmission, the angle ξ between the reference plane containing the direction of propagation and the

local vertical is a complicated function of position, but the following general points can be observed. When the electric field is parallel to the ground (horizontal), $\tau = 0$, the second term on the right-hand side of the equation for U contributes a +15-dB amount to the XPD, whereas with circular polarization the contribution is only about +0.13 dB. With the electric field vector in the reference plane containing the direction of propagation and the local vertical, $\tau = 90° - \theta$ (all angles in degrees), and the $\cos 4\tau$ term becomes $\cos 4\theta$.

5.7 Ice Depolarization

As shown in Fig. 4.3 an ice layer is present at the top of a rain region, and as noted in Table 4.1, the ice crystals can result in depolarization. The experimental evidence suggests that the chief mechanism producing depolarization in ice is differential phase shift, with little differential attenuation present. This is so because ice is a good dielectric, unlike water, which has considerable losses. Ice crystals tend to be needle-shaped or platelike and, if randomly oriented, have little effect, but depolarization occurs when they become aligned. Sudden increases in XPD that coincide with lightning flashes are thought to be a result of the lightning producing alignment. An *International Radio Consultative Committee* (CCIR) recommendation for taking ice depolarization into account is to add a fixed decibel value to the XPD value calculated for rain. Values of 2 dB are suggested for North America and 4 to 5 dB for maritime regions, and it is further suggested that the effects of ice can be ignored for time percentages less than 0.1 percent (Ippolito, 1986).

5.8 Problems and Exercises

5.1. Explain what is meant by a plane TEM wave.

5.2. Two electric fields, in time phase and each of unity amplitude, act at right angles to one another in space. On a set of x-y axes draw the path traced by the tip of the resultant electric field vector. Given that the total power developed across a 50 Ω load is 5 W, find the peak voltage corresponding to the unity amplitude.

5.3. Two electric fields with an amplitude ratio of 3:1 and in time phase, act at right angles to one another in space. On a set of x-y axes draw the path traced by the tip of the resultant. Given that the total power developed across a 50 Ω load is 10 W, find the peak voltage corresponding to the unity amplitude.

5.4. Two electric field vectors of equal amplitude are 90° out of time phase with one another. On a set of x-y axes draw the path traced by the tip of the resultant vector.

5.5. Two electric field vectors of amplitude ratio 3:1, are 90° out of time phase with one another. On a set of x-y axes draw the path traced by the tip of the resultant vector. If the peak voltages are 3 V and 1 V determine the average power developed in a 10 Ω load.

5.6. With reference to a right-hand set of rectangular coordinates, and given that Eq. (5.4) applies to a plane TEM wave, the horizontal component being directed along the x axis and the vertical component along the y axis, determine the sense of polarization of the wave.

5.7. With $\delta = -45°$ and equal amplitude components, determine the sense of polarization of a wave represented by Eq. (5.6).

5.8. A plane TEM wave has a horizontal ($+x$ directed) component of electric field of amplitude 3 V/m and a vertical ($+y$ directed) component of electric field of amplitude 5 V/m. The horizontal component leads the vertical component by a phase angle of 20°. Determine the sense of polarization.

5.9. A plane TEM wave has a horizontal ($+x$ directed) component of electric field of amplitude 3 V/m and a vertical ($+y$ directed) component of electric field of amplitude 5 V/m. The horizontal component lags the vertical component by a phase angle of 20°. Determine the sense of polarization.

5.10. Given that the plane TEM wave of Prob. 5.8 propagates in free space, determine the magnitude of the magnetic field.

5.11. Explain what is meant by *orthogonal polarization* and the importance of this in satellite communications.

5.12. The TEM wave represented by Eq. (5.4) is received by a linearly polarized antenna. Determine the reduction in emf induced in the antenna compared to what would be obtained with polarization matching.

5.13. A plane TEM wave has a horizontal ($+x$-directed) component of electric field of amplitude 3 V/m and a vertical ($+y$-directed) component of electric field of amplitude 5 V/m. The components are in time phase with one another. Determine the angle a linearly polarized antenna must be at with reference to the x axis to receive maximum signal.

5.14. For Prob. 5.13, what would be the reduction in decibels of the received signal if the antenna is placed along the x axis?

5.15. Explain what is meant by *vertical polarization* of a satellite signal. A vertically polarized wave is transmitted from a geostationary satellite and is

received at an earth station which is west of the satellite and in the northern hemisphere. Will the wave received at the earth station be vertically polarized? Give reasons for your answer.

5.16. Explain what is meant by *horizontal polarization* of a satellite signal. A horizontally polarized wave is transmitted from a geostationary satellite and is received at an earth station which is west of the satellite and in the northern hemisphere. Will the wave received at the earth station be horizontally polarized? Give reasons for your answer.

5.17. A geostationary satellite stationed at 90°W transmits a vertically polarized wave. Determine the polarization of the resulting signal received at an earth station situated at 70°W, 45°N.

5.18. A geostationary satellite stationed at 10°E transmits a vertically polarized wave. Determine the polarization of the resulting signal received at an earth station situated at 5°E, 45°N.

5.19. Explain what is meant by *cross-polarization discrimination* and briefly describe the factors which militate against good cross-polarization discrimination.

5.20. Explain the difference between *cross-polarization discrimination* and *polarization isolation*.

5.21. A linearly polarized wave traveling through the ionosphere suffers a Faraday rotation of 9°. Calculate (*a*) the polarization loss and (*b*) the cross-polarization discrimination.

5.22. Why is Faraday rotation of no concern with circularly polarized waves?

5.23. Explain how depolarization is caused by rain.

5.24. A transmission path between an earth station and a satellite has an angle of elevation of 32° with reference to the earth. The transmission is circularly polarized at a frequency of 12 GHz. Given that rain attenuation on the path is 1 dB, calculate the cross-polarization discrimination.

5.25. Repeat Prob. 5.24 for a linearly polarized signal where the electric field vector is parallel to the earth at the earth station.

5.26. Repeat Prob. 5.24 for a linearly polarized signal where the electric field vector lies in the plane containing the direction of propagation and the local vertical at the earth station.

5.27. Repeat Prob. 5.24 for a signal frequency of 18 GHz and an attenuation of 1.5 dB.

References

CCIR Report 564-2. 1982. "Propagation Data Required for Space Telecommunication System." *15 Plenary Assembly*, Vol. IX, Part 1, Geneva.
Hogg, D. C., and T. Chu. 1975. "The Role of Rain in Satellite Communications." *Proc. IEEE*, Vol. 63, No. 9, pp. 1308–1331.
Ippolito, L. J. 1986. *Radiowave Propagation in Satellite Communications* Van Nostrand Reinhold, New York.
Maral, G., and M. Bousquet. 1998. *Satellite Communications Systems.* Wiley, New York.
Miya, K. (ed.). 1981. *Satellite Communications Technology* KDD Engineering and Consulting, Japan.

Chapter 6

Antennas

6.1 Introduction

Antennas can be broadly classified according to function—as *transmitting antennas* and *receiving antennas.* Although the requirements for each function, or mode of operation, are markedly different, a single antenna may be, and frequently is, used for transmitting and receiving signals simultaneously. Many of the properties of an antenna, such as its directional characteristics, apply equally to both modes of operation, this being a result of the *reciprocity theorem* described in Sec 6.2.

Certain forms of interference (see Chap. 13) can present particular problems for satellite systems which are not encountered in other radio systems, and minimizing these requires special attention to those features of the antenna design which control interference.

Another way in which antennas for use in satellite communications can be classified is into *earth station* antennas and *satellite* or *spacecraft* antennas. Although the general principles of antennas may apply to each type, the constraints set by the physical environment lead to quite different designs in each case.

Before looking at antennas specifically for use in satellite systems, some of the general properties and definitions for antennas will be given in this and the next few sections. As already mentioned, antennas form the link between transmitting and receiving equipment, and the space propagation path. Figure 6.1*a* shows the antenna as a radiator. The power amplifier in the transmitter is shown as generating P_TW. A feeder connects this to the antenna, and the net power reaching the antenna will be P_T minus the losses in the feeder. These losses include ohmic losses and mismatch losses. The power will be further reduced by losses in the antenna so that the power radiated, shown as P_{rad}, is less than that generated at the transmitter.

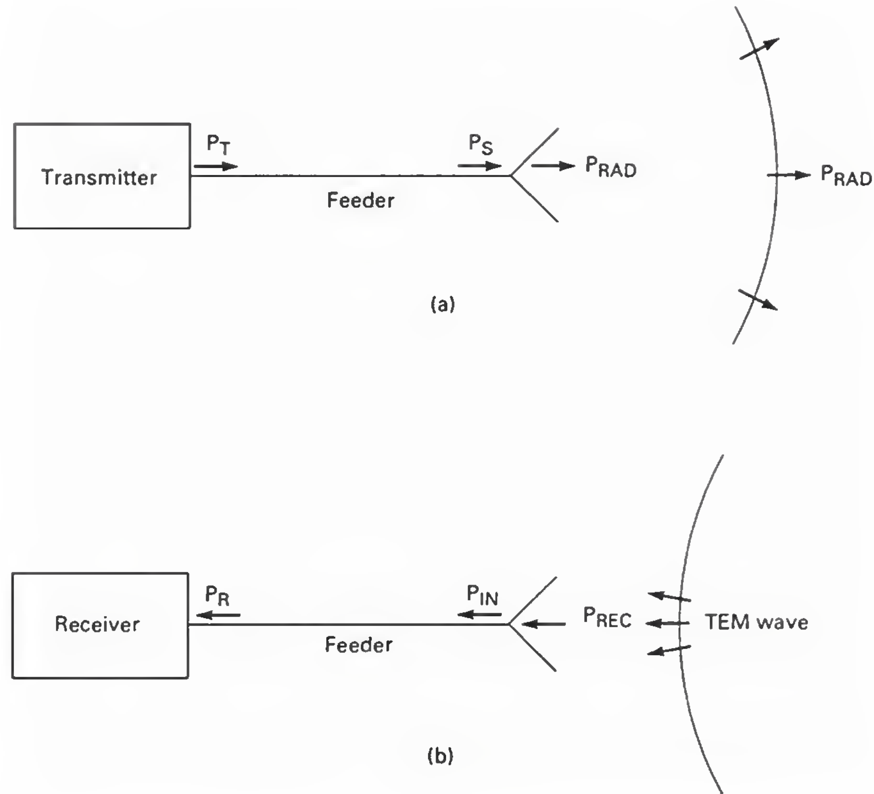


Figure 6.1 (*a*) Transmitting antenna. (*b*) Receiving antenna.

The antenna as a receiver is shown in Fig. 6.1*b*. Power P_{rec} is transferred to the antenna from a passing radio wave. Again, losses in the antenna will reduce the power available for the feeder. Receiver feeder losses will further reduce the power so that the amount P_R reaching the receiver is less than that received by the antenna.

6.2 Reciprocity Theorem for Antennas

The reciprocity theorem for antennas states that if a current I is induced in an antenna B, operated in the receive mode, by an emf applied at the terminals of antenna A operated in the transmit mode, then the same emf applied to the terminals of B will induce the same current at the terminals of A. This is illustrated in Fig. 6.2. For a proof of the reciprocity theorem, see for example, Glazier and Lamont (1958).

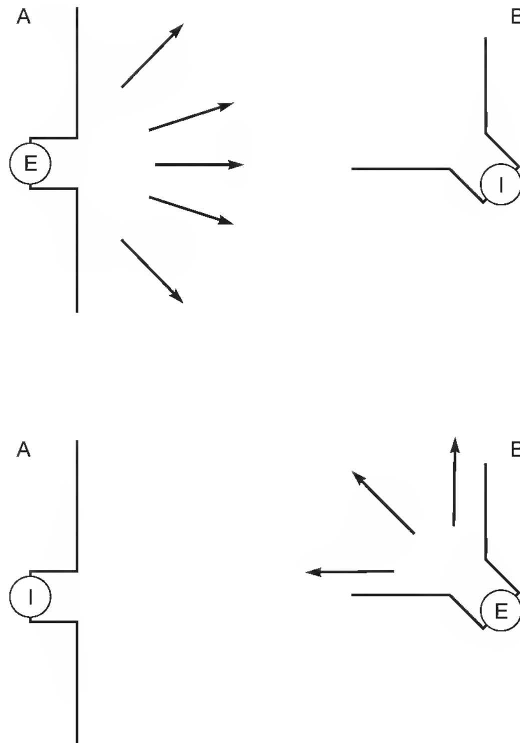


Figure 6.2 The reciprocity theorem.

A number of important consequences result from the reciprocity theorem. All practical antennas have directional patterns; that is, they transmit more energy in some directions than others, and they receive more energy when pointing in some directions than others. The reciprocity theorem requires that *the directional pattern for an antenna operating in the transmit mode is the same as that when operating in the receive mode.*

Another important consequence of the reciprocity theorem is that *the antenna impedance is the same for both modes of operation.*

6.3 Coordinate System

In order to discuss the directional patterns of an antenna, it is necessary to set up a coordinate system to which these can be referred. The system in common use is the *spherical* (or *polar*) *coordinate system* illustrated in Fig. 6.3. The antenna is imagined to be at the origin of the coordinates, and a distant point P in space is related to the origin by the coordinates r, θ, and ϕ. Thus r is the radius vector, the magnitude of

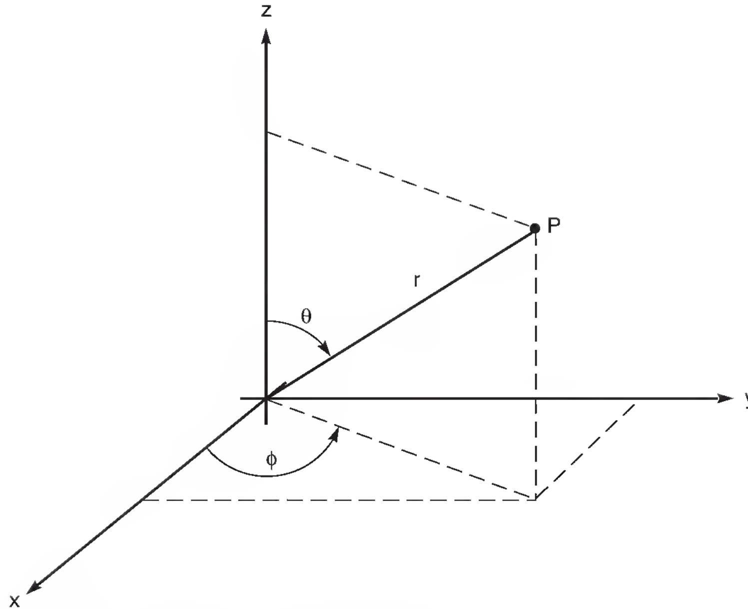


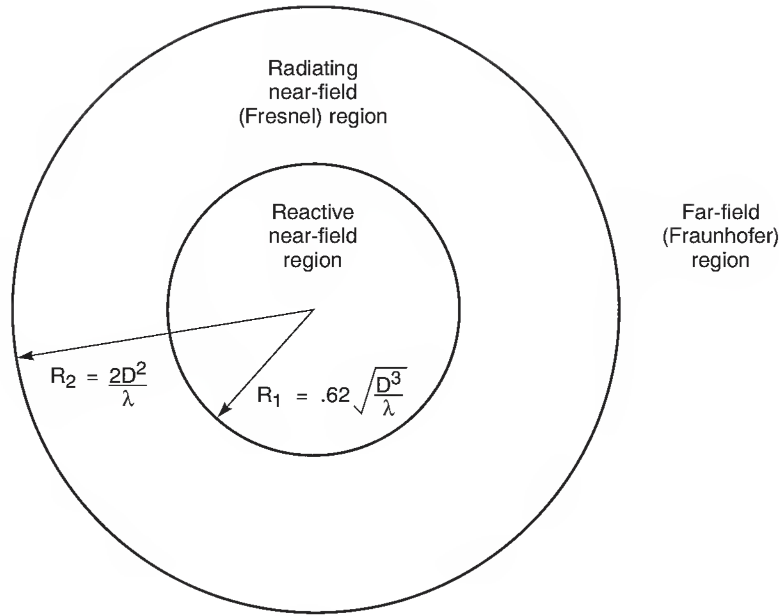
Figure 6.3 The spherical coordinate system.

which gives the distance between point P and the antenna; ϕ is the angle measured from the x axis to the projection of r in the xy plane; and θ is the angle measured from the z axis to r.

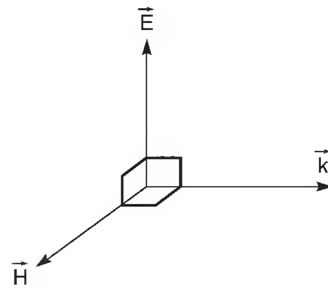
It is important to note that the x, y, and z axes form a *right-hand set.* What this means is that when one looks along the positive z direction, a clockwise rotation is required to move from the positive x axis to the positive y axis. (This is the same as the right-hand set introduced in Sec. 5.1) The right-hand set rotation becomes particularly significant when the polarization of the radio waves associated with antennas is described.

6.4 The Radiated Fields

There are three main components to the radiated electromagnetic fields surrounding an antenna: two near-field regions and a far-field region. The field strengths of the near-field components decrease rapidly with increasing distance from the antenna, one component being inversely related to distance squared, and the other to the distance cubed. At comparatively short distances these components are negligible compared with the radiated component used for radio communications, the field strength of which decreases in proportion to distance. Estimates for the distances at which the fields are significant are shown in Fig. 6.4*a*.



(a)



(b)

Figure 6.4 (*a*) The electromagnetic-field regions surrounding an antenna. (*b*) Vector diagrams in the far-field region.

Here, D is the largest dimension of the antenna (e.g., the diameter of a parabolic dish reflector), and λ is the wavelength. Only the far-field region is of interest here, which applies for distances greater than about $2D^2/\lambda$.

In the far-field region, the radiated fields form a *transverse electromagnetic* (TEM) wave in which the electric field is at right angles to the magnetic field, and both are at right angles (transverse) to the direction

of propagation. The vector relationship is shown in Fig. 6.4*b*, where **E** represents the electric field, **H** the magnetic field, and **k** the direction of propagation. These vectors form a right-hand set in the sense that when one looks along the direction of propagation, a clockwise rotation is required to go from **E** to **H.** An important practical point is that the wavefront can be assumed to be plane; that is, **E** and **H** lie in a plane to which **k** is a normal.

In the far field, the electric field vector can be resolved into two components, which are shown in relation to the coordinate system in Fig. 6.5*a*. The component labeled E_θ is tangent at point P to the circular arc of radius r. The component labeled E_ϕ is tangent at point P to the circle of radius $r \sin\theta$ centered on the z axis (this is similar to a circle of latitude on the earth's surface). Both these components are functions of θ and ϕ and in functional notation would be written as $E_\theta(\theta, \phi)$ and $E_\phi(\theta, \phi)$. The resultant magnitude of the electric field is given by

$$E = \sqrt{E_\theta^2 + E_\phi^2} \tag{6.1}$$

If E_θ and E_ϕ are peak values, E will be the peak value of the resultant, and if they are rms values, E will be the rms value of the resultant.

The vector $\mathbf{E}_0$ shown at the origin of the coordinate system represents the principal electric vector of the antenna itself. For example, for a horn antenna, this would be the electric field vector across the aperture as shown in Fig. 6.5*b*. For definiteness, the $\mathbf{E}_0$ vector is shown aligned with the y axis, since this allows two important planes to be defined:

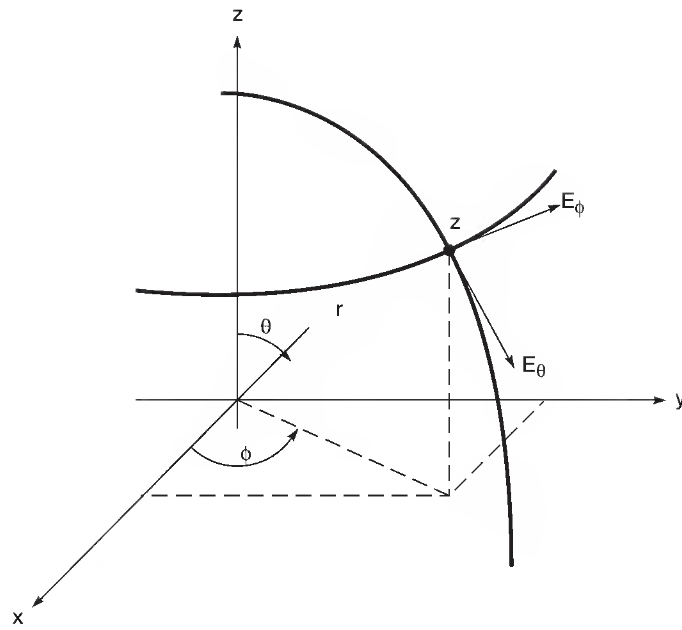
The H plane is the xz plane, for which $\phi = 0$

The E plane is the yz plane, for which $\phi = 90°$

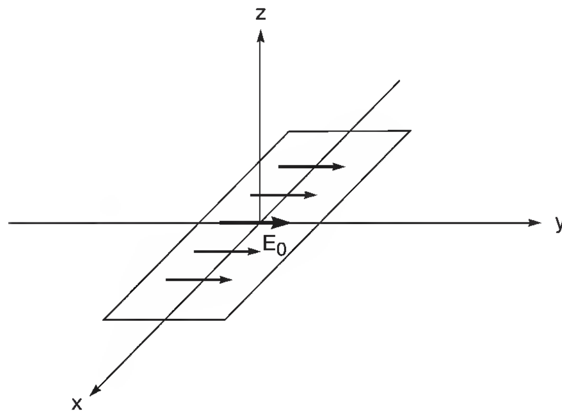
Magnetic field vectors are associated with these electric field components. Thus, following the right-hand rule, the magnetic vector associated with the E_θ component will lie parallel with E_ϕ and is normally denoted by H_ϕ, while that associated with E_ϕ will lie parallel (but pointing in the opposite direction) to E_θ and is denoted by H_θ. For clarity, the **H** fields are not shown in Fig. 6.5, but the magnitudes of the fields are related through the *wave impedance* Z_W. For radio waves in free space, the value of the wave impedance is (in terms of field magnitudes)

$$Z_W = \frac{E_\phi}{H_\theta} = \frac{E_\theta}{H_\phi} = 120\pi\Omega \tag{6.2}$$

The same value can be used with negligible error for radio waves in the earth's atmosphere.



(a)



(b)

Figure 6.5 (*a*) The electric field components E_θ and E_ϕ in the far-field region. (*b*) The reference vector E_0 at the origin.

6.5 Power Flux Density

The *power flux density* of a radio wave is a quantity used in calculating the performance of satellite communications links. The concept can be understood by imagining the transmitting antenna to be at the center of a sphere. The power from the antenna radiates outward, normal to the surface of the sphere, and the power flux density is the power flow per unit surface area. Power flux density is a vector quantity, and its magnitude is given by

$$\Psi = \frac{E^2}{Z_W} \tag{6.3}$$

Here, E is the rms value of the field given by Eq. (6.1). The units for Ψ are watts per square meter with E in volts per meter and Z_W in ohms. Because the E field is inversely proportional to distance (in this case the radius of the sphere), the power density is inversely proportional to the square of the distance.

6.6 The Isotropic Radiator and Antenna Gain

The word *isotropic* means, rather loosely, equally in all directions. Thus an *isotropic radiator* is one which radiates equally in all directions. No real antenna can radiate equally in all directions, and the isotropic radiator is therefore hypothetical. It does, however, provide a very useful theoretical standard against which real antennas can be compared. Being hypothetical, it can be made 100 percent efficient, meaning that it radiates all the power fed into it. Thus, referring back to Fig. 6.1*a*, $P_{\text{rad}} = P_S$. By imagining the isotropic radiator to be at the center of a sphere of radius r, the power flux density, which is the power flow through unit area, is

$$\Psi_i = \frac{P_S}{4\pi r^2} \tag{6.4}$$

Now the flux density from a real antenna will vary with direction, but with most antennas a well-defined maximum occurs. The *gain* of the antenna is the ratio of this maximum to that for the isotropic radiator at the same radius r:

$$G = \frac{\Psi_M}{\Psi_i} \tag{6.5}$$

A very closely related gain figure is the *directivity.* This differs from the power gain only in that in determining the isotropic flux density, the

actual power P_{rad} radiated by the real antenna is used, rather than the power P_S supplied to the antenna. These two values are related as $P_{\text{rad}} = \eta_A P_S$, where η_A is the *antenna efficiency.* Denoting the directivity by $\mathcal{D}$ gives $G = \eta_A \mathcal{D}$.

Often, the directivity is the parameter which can be calculated, and the efficiency is assumed to be equal to unity so that the power gain is also known. Note that η_A does not include feeder mismatch or polarization losses, which are accounted for separately.

The power gain G as defined by Eq. (6.5) is called the *isotropic power gain*, sometimes denoted by G_i. The power gain of an antenna also may be referred to some standard other than isotropic. For example, the gain of a reflector-type antenna may be stated relative to the antenna illuminating the reflector. Care must be taken therefore to know what reference antenna is being used when gain is stated. The isotropic gain is the most commonly used figure and will be assumed throughout this text (without use of a subscript) unless otherwise noted.

6.7 Radiation Pattern

The *radiation pattern* shows how the gain of an antenna varies with direction. Referring to Fig. 6.3, at a fixed distance r, the gain will vary with θ and ϕ and may be written generally as $G(\theta, \phi)$. The radiation pattern is the gain normalized to its maximum value. Denoting the maximum value simply by G [as given by Eq. (6.5)] the radiation pattern is

$$g(\theta, \phi) = \frac{G(\theta, \phi)}{G} \tag{6.6}$$

The radiation pattern gives the directional properties of the antenna normalized to the maximum value, in this case the maximum gain. The same function gives the power density normalized to the maximum power density. For most satellite antennas, the three-dimensional plot of the radiation pattern shows a well-defined main lobe, as sketched in Fig. 6.6*a*. In this diagram, the length of a radius line to any point on the surface of the lobe gives the value of the radiation function at that point. It will be seen that the maximum value is normalized to unity, and for convenience, this is shown pointing along the positive z axis. Be very careful to observe that the axes shown in Fig. 6.6 *do not represent distance.* The distance r is assumed to be fixed at some value in the far field. What is shown is a plot of normalized gain as a function of angles θ and ϕ.

The main lobe represents a *beam* of radiation, and the beamwidth is specified as the angle subtended by the −3-dB lines. Because in general the beam may not be symmetrical, it is usual practice to give the beamwidth in the H plane ($\phi = 0°$), as shown in Fig. 6.6*b*, and in the E plane ($\phi = 90°$), as shown in Fig. 6.6*c*.

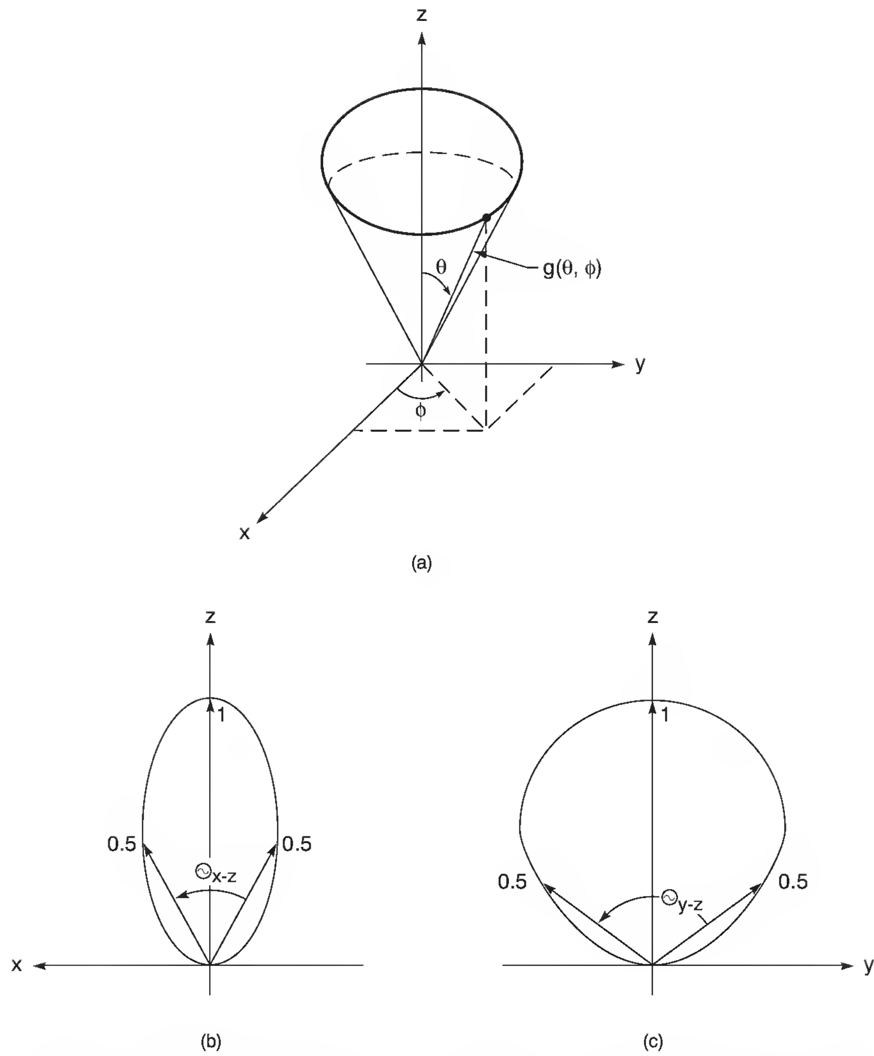


Figure 6.6 (*a*) A radiation pattern. (*b*) The beamwidth in the *H*-plane. (*c*) The beamwidth in the *E*-plane.

Because the radiation pattern is defined in terms of radiated power, the normalized electric field strength pattern will be given by $\sqrt{g(\theta, \phi)}$.

6.8 Beam Solid Angle and Directivity

Plane angles are measured in radians, and by definition, an arc of length R equal to the radius subtends an angle of one radian at the center of a circle. An angle of θ radians defines an arc length of $R\theta$ on the circle.

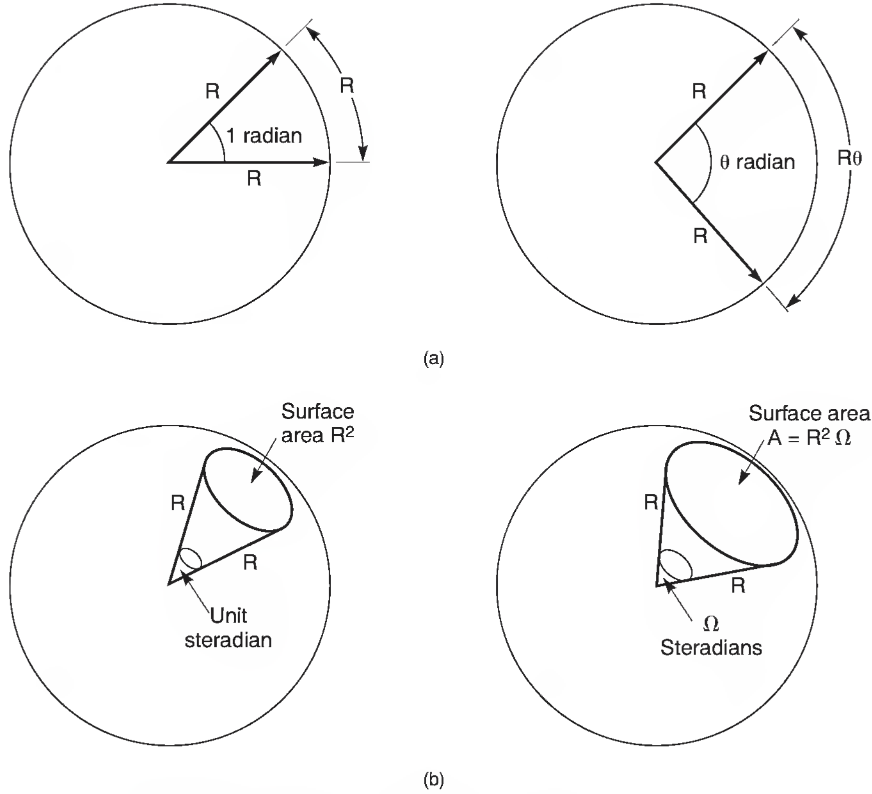


Figure 6.7 (*a*) Defining the radian. (*b*) Defining the steradian.

This is illustrated in Fig. 6.7*a*. The circumference of a circle is given by $2\pi R$, and hence the total angle subtended at the center of a circle is 2π rad. All this should be familiar to the student. What may not be so familiar is the concept of solid angle. A surface area of R^2 on the surface of a sphere of radius R subtends unit solid angle at the center of the sphere. This is shown in Fig. 6.7*b*. The unit for the solid angle is the *steradian*. A solid angle of Ω steradians defines a surface area on the sphere (a spherical cap) of $R^2\Omega$. Looking at this another way, a surface area A subtends a solid angle A/R^2 at the center of the sphere. Since the total surface area of a sphere of radius R is $4\pi R^2$, the total solid angle subtended at the center of the sphere is 4π sr.

The *radiation intensity* is the power radiated per unit solid angle. For a power P_{rad} radiated, the average radiation intensity (which is also the isotropic value) taken over a sphere is

$$U_i = \frac{P_{\text{rad}}}{4\pi} \text{ W/sr} \tag{6.7}$$

From the definition of directivity $\mathscr{D}$, the maximum radiation intensity is

$$U_{\max} = \mathscr{D}U_i \tag{6.8}$$

The *beam solid angle*, Ω_A, for an actual antenna is defined as the solid angle through which all the power would flow to produce a constant radiation intensity equal to the maximum value. Thus

$$U_{\max} = \frac{P_{\text{rad}}}{\Omega_A} \tag{6.9}$$

Combining Eqs. (6.7), (6.8), and (6.9) yields the important result

$$\mathscr{D} = \frac{4\pi}{\Omega_A} \tag{6.10}$$

This is important because for narrow-beam antennas such as used in many satellite communications systems, a good approximation to the solid angle is

$$\Omega_A \cong \text{HPBW}_E \times \text{HPBW}_H \tag{6.11}$$

where HPBW_E is the half-power beamwidth in the E plane and HPBW_H is the half-power beamwidth in the H plane, as shown in Fig. 6.6. This equation requires the half-power beamwidths to be expressed in radians, and the resulting solid angle is in steradians.

The usefulness of this relationship is that the half-power beamwidths can be measured, and hence the directivity can be found. When the half-power beamwidths are expressed in degrees, the equation for the directivity becomes

$$\mathscr{D} = \frac{41253}{\text{HPBW}^\circ_E \times \text{HPBW}^\circ_H} \tag{6.12}$$

6.9 Effective Aperture

So far, the properties of antennas have been described in terms of their radiation characteristics. A receiving antenna has directional properties also described by the radiation pattern, but in this case it refers to the ratio of received power normalized to the maximum value.

An important concept used to describe the reception properties of an antenna is that of *effective aperture*. Consider a TEM wave of a given power density Ψ at the receiving antenna. Let the load at the *antenna terminals* be a complex conjugate match so that maximum power transfer occurs and power P_{rec} is delivered to the load. Note that the power delivered to the actual receiver may be less than this as a result of

feeder losses. With the receiving antenna aligned for maximum reception (including polarization alignment, which is described in detail later), the received power will be proportional to the power density of the incoming wave. The constant of proportionality is the effective aperture A_{eff} which is defined by the equation

$$P_{\text{rec}} = A_{\text{eff}}\Psi \tag{6.13}$$

For antennas which have easily identified physical apertures, such as horns and parabolic reflector types, the effective aperture is related in a direct way to the physical aperture. If the wave could uniformly illuminate the physical aperture, then this would be equal to the effective aperture. However, the presence of the antenna in the field of the incoming wave alters the field distribution, thereby preventing uniform illumination. The effective aperture is smaller than the physical aperture by a factor known as the *illumination efficiency*. Denoting the illumination efficiency by η_I gives

$$A_{\text{eff}} = \eta_I A_{\text{physical}} \tag{6.14}$$

The illumination efficiency is usually a specified number, and it can range between about 0.5 and 0.8. Of course, it cannot exceed unity, and a conservative value often used in calculations is 0.55.

A fundamental relationship exists between the power gain of an antenna and its effective aperture. This is

$$\frac{A_{\text{eff}}}{G} = \frac{\lambda^2}{4\pi} \tag{6.15}$$

where λ is the wavelength of the TEM wave, assumed sinusoidal (for practical purposes, this will be the wavelength of the radio wave carrier). The importance of this equation is that the gain is normally the known (measurable) quantity, but once this is known, the effective aperture is also known.

6.10 The Half-Wave Dipole

The half-wave dipole is a basic antenna type which finds limited but essential use in satellite communications. Some radiation occurs in all directions except along the dipole axis itself, and it is this near-omnidirectional property which finds use for telemetry and command signals to and from the satellite, essential during the launch phase when highly directional antennas cannot be deployed.

The half-wave dipole is shown in Fig. 6.8*a*, and its radiation pattern in the *xy* plane and in any one meridian plane in Fig. 6.8*b* and *c*. Because

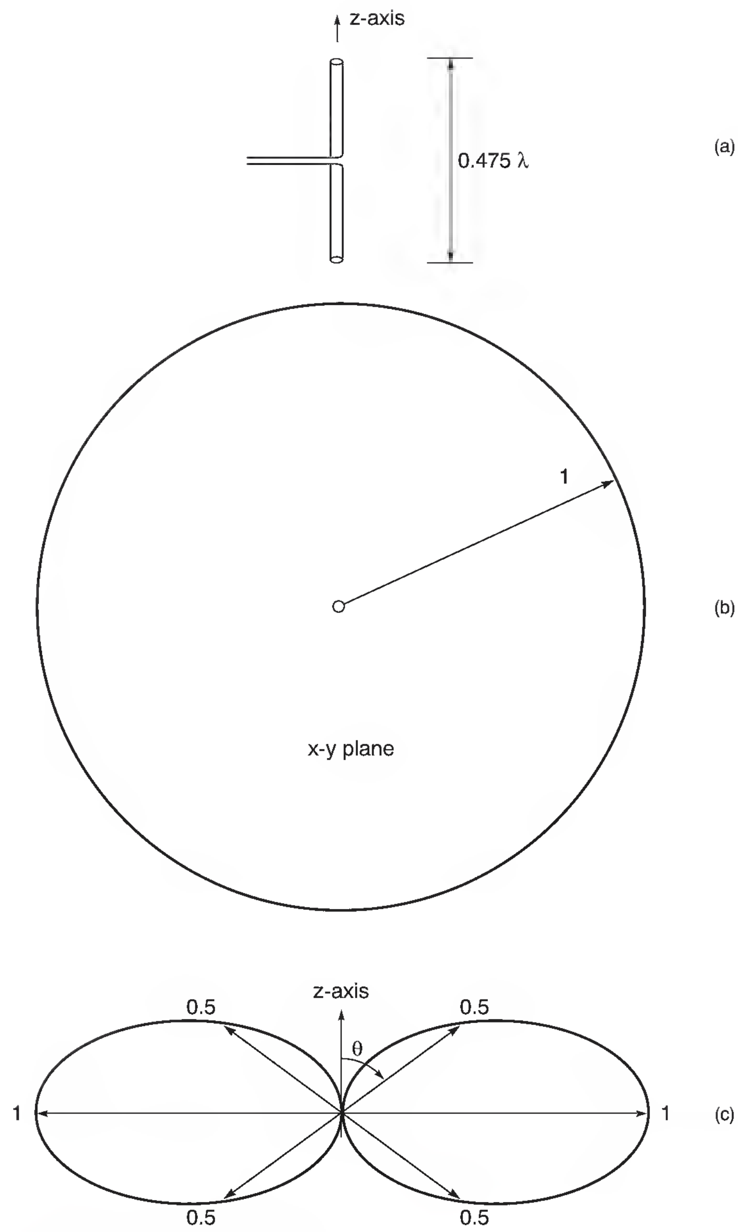


Figure 6.8 The half-wave dipole.

the phase velocity of the radio wave along the wire is somewhat less than the free-space velocity, the wavelength is also slightly less, and the antenna is cut to about 95 percent of the free-space half-wavelength. This tunes the antenna correctly to resonance. The main properties of the half-wave dipole are:

Impedance: 73 Ω
Directivity: 1.64 (or 2.15 dB)
Effective aperture: 0.13 λ^2
−3-dB beamwidth: 78°

Assuming the antenna efficiency is unit ($\eta_A = 1$), the power gain is also 1.64, or 2.15 dB. This is the gain referred to an *isotropic radiator*.

As shown in Fig. 6.8*b*, the radiation is a maximum in the *xy* plane, the normalized value being unity. The symmetry of the dipole means that the radiation pattern in this plane is a circle of unit radius. Symmetry also means that the pattern is the same for any plane containing the dipole axis (the *z* axis). Thus the radiation pattern is a function of θ only and is given by

$$g(\theta) = \frac{\cos^2\left(\frac{\pi}{2}\cos\theta\right)}{\sin^2\theta} \tag{6.16}$$

A plot of this function is shown in Fig. 6.8*c*. It is left as an exercise for the student to show that the −3-dB beamwidth obtained from this pattern is 78°.

When a satellite is launched, command and control signals must be sent and received. In the launch phase, highly directional antennas are not deployed, and a half-wave dipole, or one of its variants, is used to maintain communications.

6.11 Aperture Antennas

The open end of a waveguide is an example of a simple aperture antenna. It is capable of radiating energy being carried by the guide, and it can receive energy from a wave impinging on it. In satellite communications, the most commonly encountered aperture antennas are horn and reflector antennas. Before describing some of the practical aspects of these, the radiation pattern of an idealized aperture will be used to illustrate certain features which are important in satellite communications.

The idealized aperture is shown in Fig. 6.9. It consists of a rectangular aperture of sides *a* and *b* cut in an infinite ground plane. A uniform electric field exists across the aperture parallel to the side *b*, and the

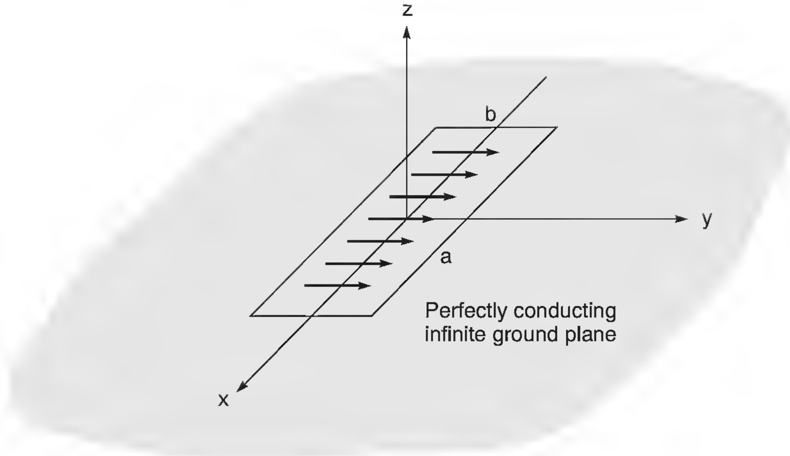


Figure 6.9 An idealized aperture radiator.

aperture is centered on the coordinate system shown in Fig. 6.3, with the electric field parallel to the y axis. Radiation from different parts of the aperture adds constructively in some directions and destructively in others, with the result that the radiation pattern exhibits a main lobe and a number of sidelobes. Mathematically, this is shown as follows:

At some fixed distance r in the far-field region, the electric field components described in Sec. 6.4 are given by

$$E_\theta(\theta, \phi) = C \sin\phi \frac{\sin X}{X} \frac{\sin Y}{Y} \tag{6.17}$$

$$E_\phi(\theta, \phi) = C \cos\theta \cos\phi \frac{\sin X}{X} \frac{\sin Y}{Y} \tag{6.18}$$

Here, C is a constant which depends on the distance r, the lengths a and b, the wavelength λ, and the electric field strength E_0. For present purposes, it can be set equal to unity. X and Y are variables given by

$$X = \frac{\pi a}{\lambda} \sin\theta \cos\phi \tag{6.19}$$

$$Y = \frac{\pi b}{\lambda} \sin\theta \sin\phi \tag{6.20}$$

It will be seen that even for the idealized and hence simplified aperture situation, the electric field equations are quite complicated. The two principal planes of the coordinate system are defined as the H plane,

which is the xz plane, for which $\phi = 0$, and the E plane, which is the xy plane, for which $\phi = 90°$. It simplifies matters to examine the radiation pattern in these two planes. Consider first the H plane. With $\phi = 0$, it is seen that $Y = 0$, $E_\theta = 0$, and

$$X = \frac{\pi a}{\lambda}\sin\theta \tag{6.21}$$

and with C set equal to unity,

$$E_\phi(\theta) = \cos\theta\frac{\sin X}{X} \tag{6.22}$$

The radiation pattern is given by:

$$\begin{aligned} g_H(\theta) &= |E_\phi(\theta)|^2 \\ &= \cos^2\theta\left|\frac{\sin X}{X}\right|^2 \end{aligned} \tag{6.23}$$

A similar analysis may be applied to the E plane resulting in $X = 0$, $E_\phi = 0$, and

$$Y = \frac{\pi b}{\lambda}\sin\theta \tag{6.24}$$

$$E_\theta(\theta) = \frac{\sin Y}{Y} \tag{6.25}$$

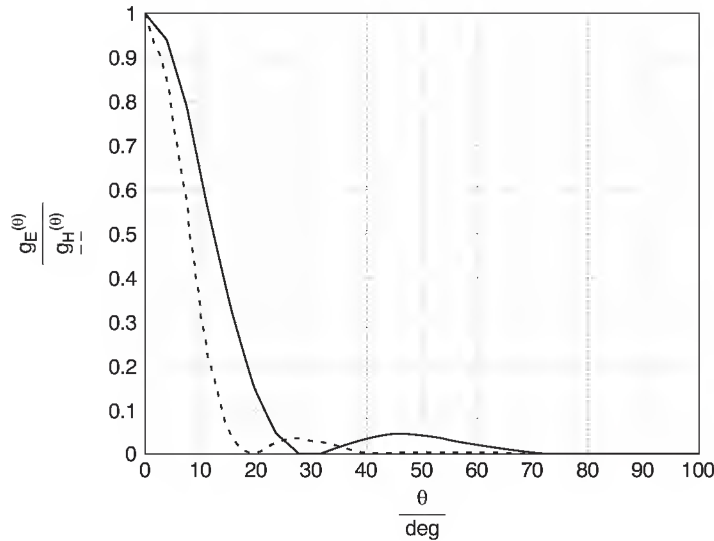
$$\begin{aligned} g_E(\theta) &= |E_\theta(\theta)|^2 \\ &= \left|\frac{\sin Y}{Y}\right|^2 \end{aligned} \tag{6.26}$$

A function that occurs frequently in communications engineering is the sampling function defined as $\mathrm{Sa}(x) = \sin x/x$. This function is available in tabular form in many handbooks, and may also be available in programmable calculators. A point to bear in mind when evaluating this function is that the denominator x must be in radians. $\mathrm{Sa}(x) = 1$ for $x = 0$, and $\mathrm{Sa}(x) = 0$ for $x = n\pi$, where n is the integer. It is seen that the H plane pattern contains the function $\mathrm{Sa}(Y)$ and the E plane, the function $\mathrm{Sa}(X)$. These radiation patterns are illustrated in Example 6.1.

Example 6.1 Plot the E-plane and H-plane radiation patterns for the uniformly illuminated aperture for which $a = 3\lambda$, and $b = 2\lambda$.

Solution Looking first at the H plane, for $a = 3\lambda$, $X = 3\pi\sin\theta$. As noted here, the sampling function has well-defined zeros, occurring in this case when $\sin\theta = 1/3$, 2/3, or 1. The $g_H(\theta)$ function will have correspondingly zeros or nulls. (The $\cos\theta$ term will also have zeros for $\theta = n\pi/2$, where n is any odd integer.

For the E plane and $b = 2\lambda$, $Y = 2\pi\sin\theta$. Again, the sampling function has well-defined zeros occurring in this case when $\sin\theta = 1/2$ or 1, and the $g_E(\theta)$ function



shows corresponding zeros, or nulls. Plots of the radiation functions are shown. The curves will be symmetrical about the vertical axis and so only one-half of the curves need be shown.

The results of Example 6.1 show the main lobe and the sidelobes. These are a general feature of aperture antennas. The sidelobes can result in interference to adjacent channels, and maximum allowable levels are specified to minimize this, (see Fig. 6.20). The nulls in the radiation pattern can be useful in some situations where these can be aligned with an interfering source.

The uniform field distribution across the aperture (Fig. 6.9) cannot be realized in practice, the actual distribution depending on the manner in which the aperture is energized. In practice, therefore, the radiation pattern will depend on the way the aperture is energized. It is also influenced by the physical construction of the antenna. With reflector-type antennas, for example, the position of the primary feed can change the pattern in important ways.

Another important practical consideration with real antennas is the *cross-polarization* which can occur. This refers to the antenna in the transmit mode radiating, and in the receive mode responding to, an unwanted signal with polarization orthogonal to the desired polarization (see Sec. 5.2). As mentioned in Chap. 5, frequency reuse makes use of orthogonal polarization, and any unwanted cross-polarized component results in interference. The cross-polarization characteristics of some practical antennas will be looked at in the following sections.

The aperture shown in Fig. 6.9 is linearly polarized, the **E** vector being directed along the y axis. At some arbitrary point in the far-field region, the wave will remain linearly polarized, the magnitude E being given by Eq. (6.1). It is only necessary for the receiving antenna to be oriented so that E induces maximum signal, with no component orthogonal to E so that cross-polarization is absent. Care must be taken, however, in how cross-polarization is defined. The linearly polarized field **E** can be resolved into two vectors, one parallel to the plane containing the aperture vector E_0, referred to as the *copolar* component, and a second component orthogonal to this, referred to as the cross-polarized component. The way in which these components are used in antenna measurements is detailed in Chang (1989) and Rudge et al. (1982).

6.12 Horn Antennas

The horn antenna is an example of an aperture antenna that provides a smooth transition from a waveguide to a larger aperture that couples more effectively into space. Horn antennas are used directly as radiators aboard satellites to illuminate comparatively large areas of the earth, and they are also widely used as primary feeds for reflector type antennas both in transmitting and receiving modes. The three most commonly used types of horns are illustrated in Fig. 6.10.

6.12.1 Conical horn antennas

The *smooth-walled* conical antenna shown in Fig. 6.10 is the simplest horn structure. The term smooth-walled refers to the inside wall. The horn may be fed from a rectangular waveguide, but this requires a rectangular-to-circular transition at the junction. Feeding from a circular guide is

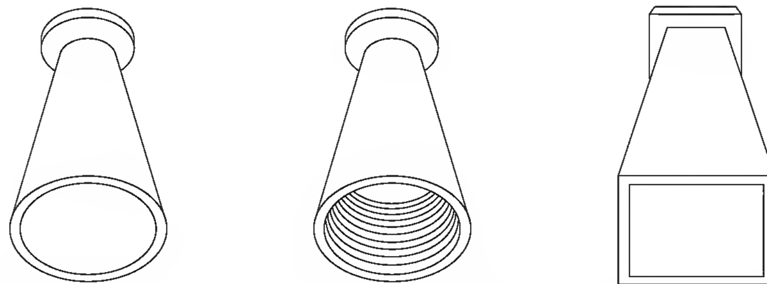


Figure 6.10 Horn antennas: (*a*) smooth-walled conical, (*b*) corrugated, and (*c*) pyramidal.

direct and is the preferred method, with the guide operating in the TE_{11} mode. The conical horn antenna may be used with linear or circular polarization, but in order to illustrate some of the important features, linear polarization will be assumed.

The electric field distribution at the horn mouth is sketched in Fig. 6.11 for vertical polarization. The curved field lines can be resolved into vertical and horizontal components as shown. The TEM wave in the far field is linearly polarized, but the horizontal components of the aperture field give rise to cross-polarized waves in the far-field region. Because of the symmetry, the cross-polarized waves cancel in the principal planes (the E and H planes); however, they produce four peaks, one in each quadrant around the main lobe. Referring to Fig. 6.5, the cross-polarized fields peak in the $\phi = \pm 45°$ planes. The peaks are about -19 dB relative to the peak of the main (copolar) lobe (Olver, 1992).

The smooth-walled horn does not produce a symmetrical main beam, even though the horn itself is symmetrical. The radiation patterns are complicated functions of the horn dimensions. Details will be found in Chang (1989), where it is shown that the beamwidths in the principal planes can differ widely. This lack of symmetry is a disadvantage where global coverage is required.

By operating a conical horn in what is termed a *hybrid mode*, which is a nonlinear combination of transverse electric (TE) and transverse magnetic (TM) modes, the pattern symmetry is improved, the cross-polarization is reduced, and a more efficient main beam is produced with low sidelobes. It is especially important to reduce the cross-polarization where frequency reuse is employed, as described in Sec. 5.2.

One method of achieving a hybrid mode is to corrugate the inside wall of the horn, thus giving rise to the *corrugated horn* antenna. The cross section of a corrugated horn is shown in Fig. 6.12*a*. The aperture electric field is shown in Fig. 6.12*b*, where it is seen to have a much lower cross-polarized component. This field distribution is sometimes referred to as a *scalar field* and the horn as a *scalar horn*. A development of the scalar horn is the scalar feed, Fig. 6.13, which can be seen on most

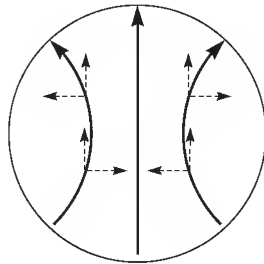


Figure 6.11 Aperture field in a smooth-walled conical horn.

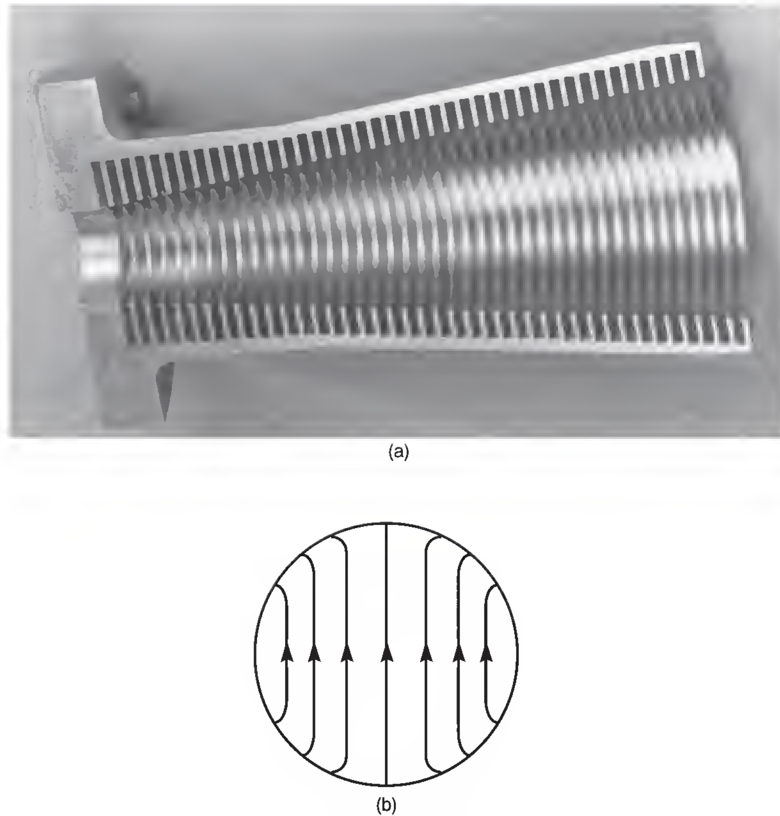


Figure 6.12 (*a*) Cross section of a corrugated horn. (*Courtesy of Alver, 1992.*) (*b*) Aperture field.

domestic receiving systems. Here, the flare angle of the horn is 90°, and the corrugations are in the form of a flange surrounding the circular waveguide. The corrugated horn is obviously more difficult to make than the smooth-walled version, and close manufacturing tolerances must be maintained, especially in machining the slots or corrugations, all of which contribute to increased costs. A comprehensive description of the corrugated horn will be found in Olver (1992), and design details will be found in Chang (1989).

A hybrid mode also can be created by including a dielectric rod along the axis of the smooth-walled horn, this being referred to as a *dielectric-rod-loaded antenna* (see Miya, 1981).

A *multimode* horn is one which is excited by a linear combination of transverse electric and transverse magnetic fields, the most common type being the *dual-mode horn*, which combines the TE_{11} and TM_{11} modes. The advantages of the dual-mode horn are similar to those of



Figure 6.13 A scalar feed.

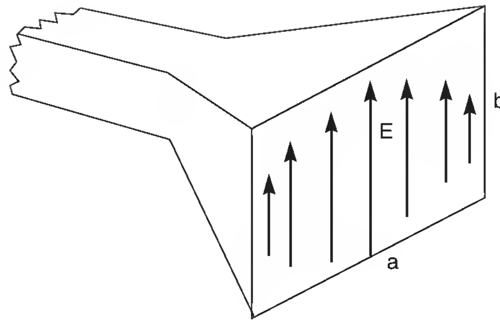
the hybrid-mode horn, that is, better main lobe symmetry, lower cross-polarization, and a more efficient main beam with low sidelobes. Dual-mode horns have been installed aboard various satellites (see Miya, 1981).

Horns which are required to provide earth coverage from geostationary satellites must maintain low cross-polarization and high gain over a cone angle of $\pm 9°$. This is achieved more simply and economically with dual-mode horns (Hwang, 1992).

6.12.2 Pyramidal horn antennas

The pyramidal horn antenna, illustrated in Fig. 6.14, is primarily designed for linear polarization. In general, it has a rectangular cross section $a \times b$ and operates in the TE_{10} waveguide mode, which has the electric field distribution shown in Fig. 6.14.

In general, the beamwidths for the pyramidal horn differ in the E and H planes, but it is possible to choose the aperture dimensions to make these equal. The pyramidal horn can be operated in horizontally and vertically polarized modes simultaneously, giving rise to dual-linear polarization. According to Chang (1989), the cross-polarization characteristics of the pyramidal horn have not been studied to a great extent, and if required, they should be measured.



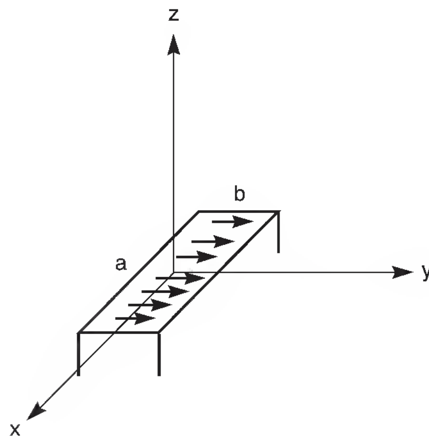


Figure 6.14 The pyramidal horn.

For any of the aperture antennas discussed, the isotropic gain can be found in terms of the area of the physical aperture by using the relationships given in Eqs. (6.14) and (6.15). For accurate gain determinations, the difficulties lie in determining the illumination efficiency η_I, which can range from 35 to 80 percent for horns and from 50 to 80 percent for circular reflectors (Balanis, 1982, p. 475). Circular reflectors are discussed in Sec. 6.13.

6.13 The Parabolic Reflector

Parabolic reflectors are widely used in satellite communications systems to enhance the gain of antennas. The reflector provides a focusing mechanism which concentrates the energy in a given direction. The most

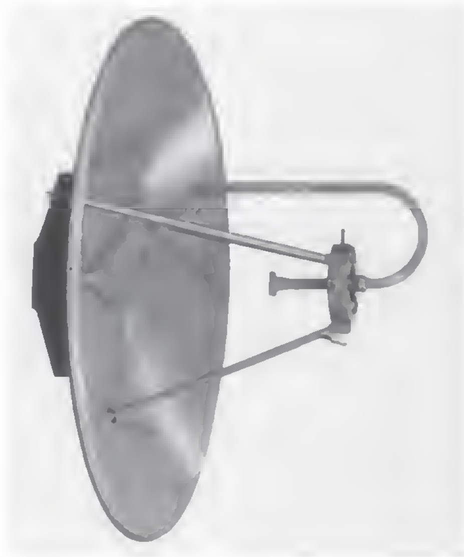


Figure 6.15 A parabolic reflector. (*Courtesy of Scientific Atlanta, Inc.*)

commonly used form of parabolic reflector has a circular aperture, as shown in Fig. 6.15. This is the type seen in many home installations for the reception of TV signals. The circular aperture configuration is referred to as a *paraboloidal reflector*.

The main property of the paraboloidal reflector is its focusing property, normally associated with light, where parallel rays striking the reflector converge on a single point known as the *focus* and, conversely, rays originating at the focus are reflected as a parallel beam of light. This is illustrated in Fig. 6.16. Light, of course, is a particular example of an electromagnetic wave, and the same properties apply to electromagnetic waves in general, including the radio waves used in satellite communications. The ray paths from the focus to the aperture plane (the plane containing the circular aperture) are all equal in length.

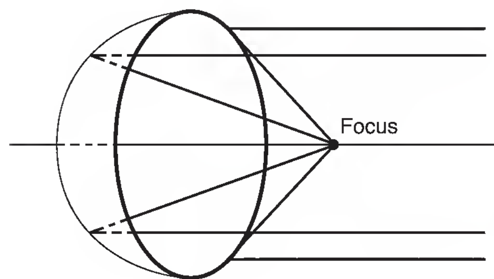


Figure 6.16 The focusing property of a paraboloidal reflector.

The geometric properties of the paraboloidal reflector of interest here are most easily demonstrated by means of the *parabola*, which is the curve traced by the reflector on any plane normal to the aperture plane and containing the focus. This is shown in Fig. 6.17*a*. The *focal point* or *focus* is shown as *S*, the *vertex* as *A*, and the axis is the line passing through *S* and *A*. SP is the *focal distance* for any point *P* and SA the *focal length*, usually denoted by *f*. (The parabola is examined in more detail in App. B). A ray path is shown as SPQ, where *P* is a point on the curve and *Q* is a point in the aperture plane. Length PQ lies parallel to the axis. For any point *P*, all path lengths SPQ are equal; that is, the distance SP + PQ is a constant which applies for all such paths. The path equality means that a wave originating from an isotropic point source has a uniform phase distribution over the aperture plane. This property, along with the parallel-beam property, means that the wavefront is plane. Radiation from the paraboloidal reflector appears to originate as a plane wave from the plane, normal to the axis

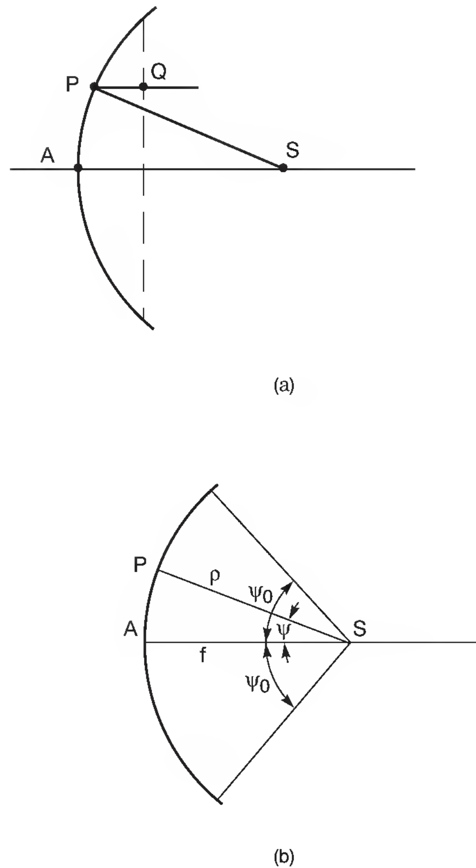


Figure 6.17 (*a*) The focal length f = SA and a ray path SPQ. (*b*) The focal distance ρ.

and containing the directrix (see App. B). Although the characteristics of the reflector antenna are more readily described in terms of radiation, it should be kept in mind that the reciprocity theorem makes these applicable to the receiving mode as well.

Now although there are near- and far-field components present in the reflector region, the radio link is made through the far-field component, and only this need be considered. For this, the reflected wave is a plane wave, while the wave originating from the isotropic source and striking the reflector has a spherical wavefront. The power density in the plane wave is independent of distance. However, for the spherical wave reaching the reflector from the source, the power density of the far-field component decreases in inverse proportion to the distance squared, and therefore, the illumination at the edge of the reflector will be less than that at the vertex. This gives rise to a nonuniform amplitude distribution across the aperture plane, which in effect means that the illumination efficiency is reduced. Denoting the focal distance by ρ and the focal length by f as in Fig. 6.17*b*, then, as shown in App. B,

$$\frac{\rho}{f} = \sec^2 \frac{\Psi}{2} \tag{6.27}$$

The *space attenuation function* (SAF) is the ratio of the power reaching point P to that reaching point A, and since the power density is inversely proportional to the square of the distance, the ratio is given by

$$\begin{aligned} \text{SAF} &= \left(\frac{f}{\rho}\right)^2 \\ &= \cos^4 \frac{\Psi}{2} \end{aligned} \tag{6.28}$$

For satellite applications, a high illumination efficiency is desirable. This requires that the radiation pattern of the primary antenna, which is situated at the focus and which illuminates the reflector, should approximate as closely as practical the inverse of the space attenuation factor.

An important ratio is that of aperture diameter to focal length. Denoting the diameter by D, then, as shown in App. B,

$$\frac{f}{D} = 0.25 \cot \frac{\Psi_0}{2} \tag{6.29}$$

The position of the focus in relation to the reflector for various values of f/D is shown in Fig. 6.18. For $f/D < 0.25$, the primary antenna lies in the space between the reflector and the aperture plane, and the illumination tapers away toward the edge of the reflector. For $f/D > 0.25$, the primary antenna lies outside the aperture plane, which results in more

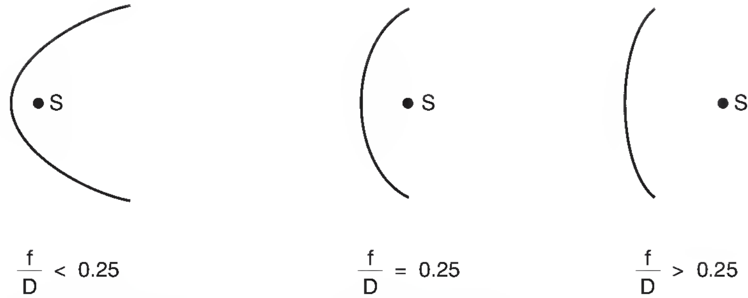


Figure 6.18 Position of the focus for various *f/D* values.

nearly uniform illumination, but *spillover* increases. In the transmitting mode, spillover is the radiation from the primary antenna which is directed toward the reflector but which lies outside the angle $2\Psi_0$. In satellite applications, the primary antenna is usually a horn (or an array of horns, as will be shown later) pointed toward the reflector. In order to compensate for the space attenuation described earlier, higher-order modes can be added to the horn feed so that the horn-radiation pattern approximates the inverse of the space attenuation function (Chang, 1989).

The radiation from the horn will be a spherical wave, and the *phase center* will be the center of curvature of the wavefront. When used as the primary antenna for a parabolic reflector, the horn is positioned so that the phase center lies on the focus.

The focal length can be given in terms of the depth of the reflector and its diameter. It is sometimes useful to know the focal length for setting up a receiving system. The depth *d* is the perpendicular distance from the aperture plane to the vertex. This relationship is shown in App. B [Eq. (B.37)] to be

$$f = \frac{D^2}{16d} \tag{6.30}$$

The gain and beamwidths of the paraboloidal antenna are as follows. The physical area of the aperture plane is

$$\text{Area} = \frac{\pi D^2}{4} \tag{6.31}$$

From the relationships given by Eqs. (6.14) and (6.15), the gain is

$$\begin{aligned} G &= \frac{4\pi}{\lambda^2}\eta_I \text{area} \\ &= \eta_I \left(\frac{\pi D}{\lambda}\right)^2 \end{aligned} \tag{6.32}$$

The radiation pattern for the paraboloidal reflector is similar to that developed in Example 6.1 for the rectangular aperture, in that there is a main lobe and a number of sidelobes, although there will be differences in detail. In practice, the sidelobes are accounted for by an envelope function as described in Sec. 13.2.4. Useful approximate formulas for the half-power beamwidth and the beamwidth *between the first nulls* (BWFN) are

$$\text{HPBW} \cong 70\frac{\lambda}{D} \tag{6.33}$$

$$\text{BWFN} \cong 2\text{HPBW} \tag{6.34}$$

In these relationships, the beamwidths are given in degrees. The paraboloidal antenna described so far is *center-fed*, in that the primary horn is pointed toward the center of the reflector. With this arrangement the primary horn and its supports present a partial blockage to the reflected wave. The energy scattered by the blockage is lost from the main lobe, and it can create additional sidelobes. One solution is to use an *offset feed* as described in Sec. 6.14.

The wave from the primary radiator induces surface currents in the reflector. The curvature of the reflector causes the currents to follow curved paths so that both horizontal and vertical components are present, even where the incident wave is linearly polarized in one or other of these directions. The situation is sketched for the case of vertical polarization in Fig. 6.19. The resulting radiation consists of copolarized and cross-polarized fields. The symmetry of the arrangement means that the cross-polarized component is zero in the principal planes (the E and H planes). Cross-polarization peaks in the $\phi = \pm 45°$ planes, assuming a coordinate system as shown in Fig. 6.5a. Sketches of the copolar and cross-polar radiation patterns for the 45° planes are shown in Fig. 6.20.

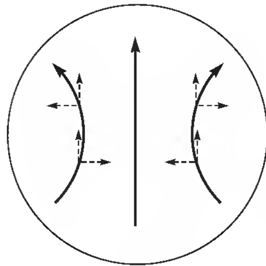


Figure 6.19 Current paths in a paraboloidal reflector for linear polarization.

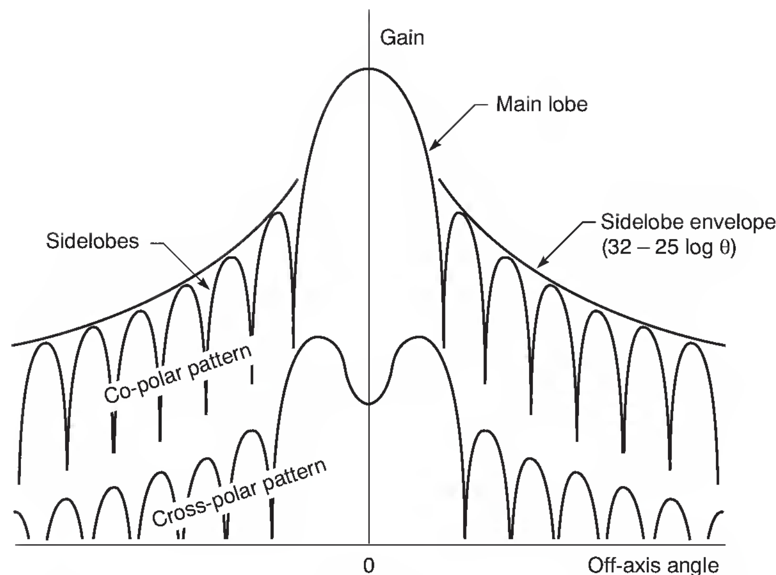
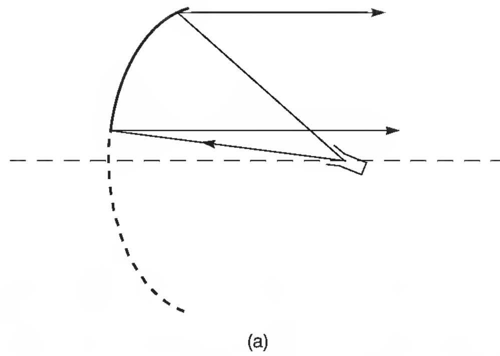


Figure 6.20 Copolar and cross-polar radiation patterns. (*Courtesy of FCC Report FCC/OST R83-2, 1983.*)

6.14 The Offset Feed

Figure 6.21*a* shows a paraboloidal reflector with a horn feed at the focus. In this instance the radiation pattern of the horn is offset so that it illuminates only the upper portion of the reflector. The feed horn and its support can be placed well clear of the main beam so that no blockage occurs. With the center-fed arrangement described in the previous section, the blockage results typically in a 10 percent reduction in efficiency (Brain and Rudge, 1984) and increased radiation in the sidelobes. The offset arrangement avoids this. Figure 6.21*b* shows a development model of an offset antenna intended for use in the European Olympus satellite.

The main disadvantages of the offset feed are that a stronger mechanical support is required to maintain the reflector shape, and because of the asymmetry, the cross-polarization with a linear polarized feed is worse compared with the center-fed antenna. Polarization compensation can be introduced into the primary feed to correct for the cross-polarization, or a *polarization-purifying grid* can be incorporated into the antenna structure (Brain and Rudge, 1984). The advantages of the offset feed are sufficiently attractive for it to be standard on many satellites (see, e.g., Figs. 7.6 and 7.22). It is also used with double-reflector earth station antennas, as shown in Fig. 6.24 later, and is being used increasingly with small receive-only earth station antennas.



(a)



(b)

Figure 6.21 (*a*) Ray paths for an offset reflector. (*b*) The offset feed for a paraboloidal reflector. (*Courtesy of Brain and Rudge, 1984.*)

6.15 Double-Reflector Antennas

With reflector-type antennas, the feeder connecting the feed horn to the transmit/receive equipment must be kept as short as possible to minimize losses. This is particularly important with large earth stations where the transmit power is large and where very low receiver noise is required. The single-reflector system described in Sec. 6.14 does not lend itself very well to achieving this, and more satisfactory, but more costly, arrangements are possible with a double-reflector system. The feed horn is mounted at the rear of the main reflector through an opening at the vertex, as illustrated in Fig. 6.22. The rear mount makes for a compact feed, which is an advantage where steerable antennas must be used, and access for servicing is easier. The subreflector, which is mounted at the front of the main reflector, is generally smaller than the feed horn and causes less blockage. Two main types are in use, the Cassegrain antenna and the Gregorian antenna, named after the astronomers who first developed them.

6.15.1 Cassegrain antenna

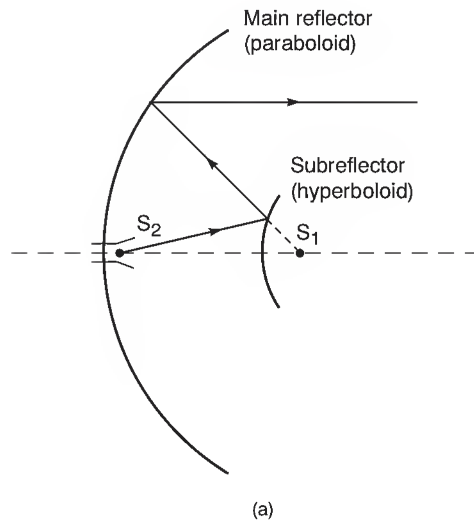
The basic Cassegrain form consists of a main paraboloid and a subreflector, which is a hyperboloid (see App. B). The subreflector has two



Figure 6.22 A 19-m Cassegrain antenna. (*Courtesy of TIW Systems, Inc.*)

focal points, one of which is made to coincide with that of the main reflector and the other with the phase center of the feed horn, as shown in Fig. 6.23*a*. The Cassegrain system is equivalent to a single paraboloidal reflector of focal length

$$f_e = \frac{e_h + 1}{e_h - 1} f \tag{6.35}$$



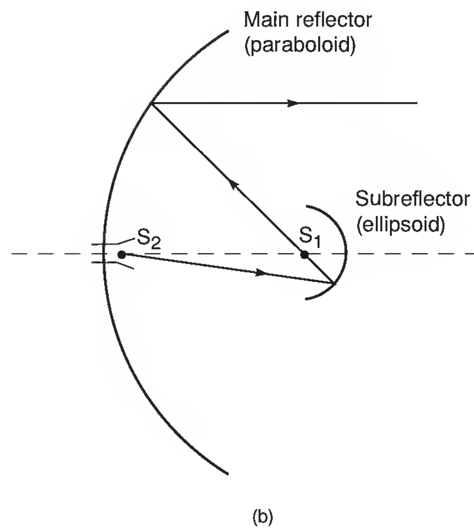


Figure 6.23 Ray paths for (*a*) Cassegrain antenna. (*b*) Gregorian antenna.

where e_h is the eccentricity of the hyperboloid (see App. B) and f is the focal length of the main reflector. The eccentricity of the hyperboloid is always greater than unity and typically ranges from about 1.4 to 3. The equivalent focal length, therefore, is greater than the focal length of the main reflector. The diameter of the equivalent paraboloid is the same as that of the main reflector, and hence the f/D ratio is increased. As shown in Fig. 6.18, a large f/D ratio leads to more uniform illumination, and in the case of the Cassegrain, this is achieved without the spillover associated with the single-reflector system. The larger f/D ratio also results in lower cross-polarization (Miya, 1981). The Cassegrain system is widely used in large earth-station installations.

6.15.2 Gregorian antenna

The basic Gregorian form consists of a main paraboloid and a subreflector, which is an ellipsoid (see App. B). As with the hyperboloid, the subreflector has two focal points, one of which is made to coincide with that of the main reflector and the other with the phase center of the feed horn, as shown in Fig. 6.23*b*. The performance of the Gregorian system is similar in many respects to the Cassegrain. An offset Gregorian antenna is illustrated in Fig. 6.24.

6.16 Shaped Reflector Systems

With the double-reflector systems described, the illumination efficiency of the main reflector can be increased while avoiding the problem of increased spillover by shaping the surfaces of the subreflector and main reflector. With the Cassegrain system, for example, altering the curvature of the center section of the subreflector to be greater than that of the hyperboloid allows it to reflect more energy toward the edge of the main reflector, which makes the amplitude distribution more uniform. At the same time, the curvature of the center section of the main reflector is made smaller than that required for the paraboloid. This compensates for the reduced path length so that the constant phase condition across the aperture is maintained. The edge of the subreflector surface is shaped in a manner to reduce spillover, and of course, the overall design must take into account the radiation pattern of the primary feed. The process, referred to as *reflector shaping*, employs computer-aided design methods. Further details will be found in Miya (1981) and Rusch (1992).

With the Hughes shaped reflector (Fig. 6.25), dimples and/or ripples are created on the surface. The depth of these is no more than a wavelength, which makes them rather difficult to see, especially at the Ka band. Reflections from the uneven surface reinforce radiation in some

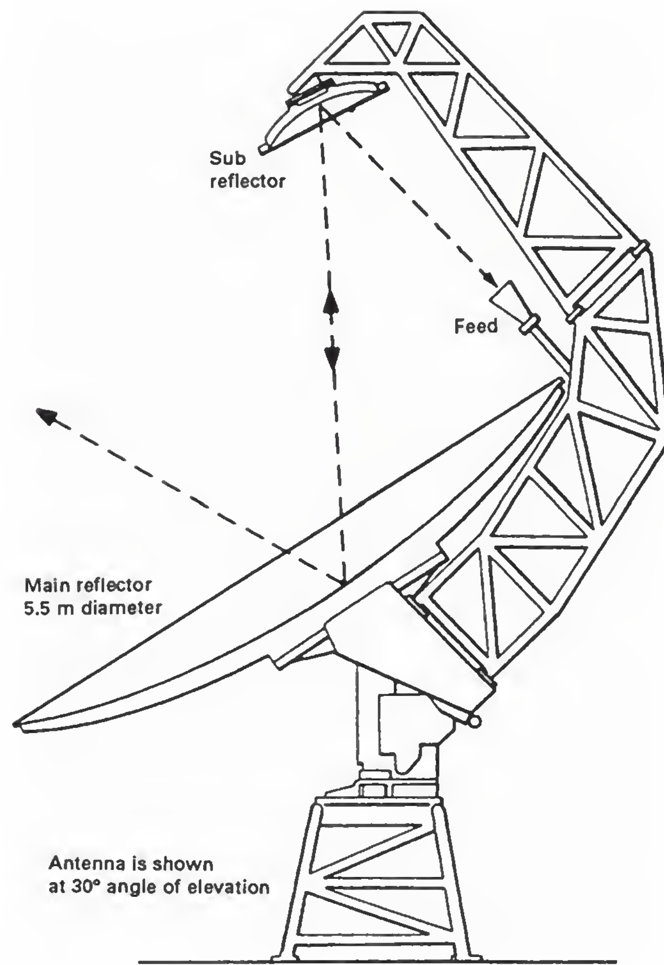


Figure 6.24 Offset Gregorian antenna. (*Courtesy of Radio Electr. Eng., vol. 54, No. 3, Mar. 1984, p. 112.*)

directions and reduce it in others. The design steps start with a map of the ground coverage area desired. A grid is overlaid on the map, and at each grid intersection a weighting factor is assigned which corresponds to the antenna gain desired in that direction. The intersection points on the coverage area also can be defined by the azimuth and elevation angles required at the ground stations, which enables the beam contour to be determined. The beam-shaping stage starts by selecting a smooth parabolic reflector that forms an elliptical beam encompassing the coverage area. The reflector surface is computer modeled as a series of mathematical functions that are changed or perturbed

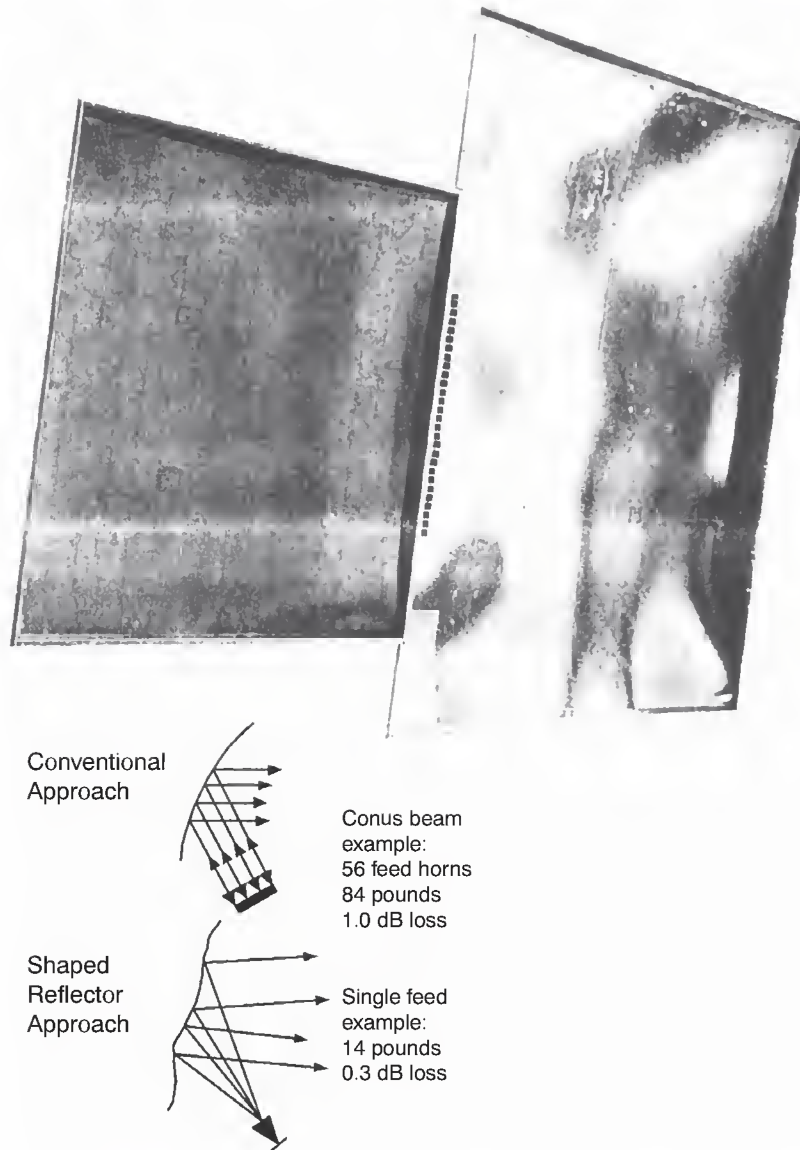


Figure 6.25 Shaped-beam reflector, showing ray paths. (*Courtesy of Hughes Space and Communications Company. Reproduced from Vectors XXXV(3):14, 1993. © Hughes Aircraft Co.*)

until the model produces the desired coverage. On a first pass the computer analyzes the perturbations and translates these into surface ripples. The beam footprint computed for the rippled surface is compared with the coverage area. The perturbation analysis is refined and the passes are repeated until a satisfactory match is obtained. As

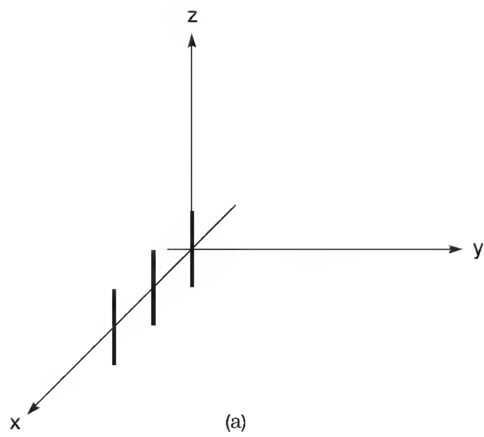
an example of the improvements obtained, the conventional approach to producing a CONUS beam requires 56 feed horns, and the feed weighs 84 pounds and has a 1-dB loss. With a shaped reflector, a single-feed horn is used, and it weighs 14 pounds and has 0.3-dB loss (see Vectors, 1993).

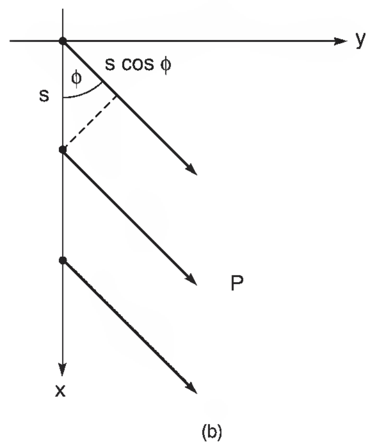
Shaped reflectors also have been used to compensate for rainfall attenuation, and this has particular application in *direct broadcast satellite* (DBS) systems (see Chap. 16). In this case, the reflector design is based on a map similar to that shown in Fig. 16.8, which gives the rainfall intensity as a function of latitude and longitude. The attenuation resulting from the rainfall is calculated as shown in Sec. 4.4, and the reflector is shaped to redistribute the radiated power to match, within practical limits, the attenuation.

6.17 Arrays

Beam shaping can be achieved by using an array of basic elements. The elements are arranged so that their radiation patterns provide mutual reinforcement in certain directions and cancellation in others. Although most arrays used in satellite communications are two-dimensional horn arrays, the principle is most easily explained with reference to an in-line array of dipoles (Fig. 6.26*a* and *b*). As shown previously (Fig. 6.8), the radiation pattern for a single dipole in the *xy* plane is circular, and it is this aspect of the radiation pattern that is altered by the array configuration. Two factors contribute to this: the difference in distance from each element to some point in the far field and the difference in the current feed to each element. For the coordinate system shown in Fig. 6.26*b*, the *xy* plane, the difference in distance is given by $s\cos\phi$. Although this distance is small compared with the range between the array and point P, it plays a crucial role in determining the phase relationships between the radiation from each element. It should be kept in mind that at any point in the far field the array appears as a point source, the situation being as sketched in Fig. 6.26*c*. For this analysis, the point P is taken to lie in the *xy* plane. Since a distance of one wavelength corresponds to a phase difference of 2π, the phase lead of element n relative to $n - 1$ resulting from the difference in distance is $(2\pi/\lambda)s\cos\phi$. To illustrate the array principles, it will be assumed that each element is fed by currents of equal magnitude but differing in phase progressively by some angle α. Positive values of α mean a phase lead and negative values a phase lag. The total phase lead of element n relative to $n - 1$ is therefore

$$\Psi = \alpha + \frac{2\pi}{\lambda}s\cos\phi \tag{6.36}$$





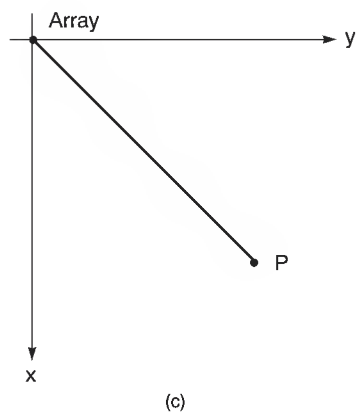


Figure 6.26 An in-line array of dipoles.

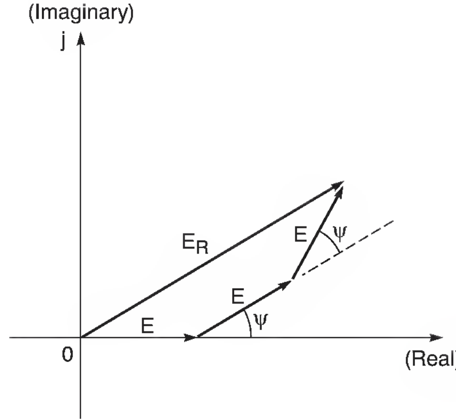


Figure 6.27 Phasor diagram for the in-line array of dipoles.

The Argand diagram for the phasors is shown in Fig. 6.27. The magnitude of the resultant phasor can be found by first resolving the individual phasors into horizontal (real axis) and vertical (imaginary axis) components, adding these, and finding the resultant. The contribution from the first element is E, and from the second element, $E\cos\Psi + jE\sin\Psi$. The third element contributes $E\cos 2\Psi + jE\sin 2\Psi$, and in general the Nth element contributes $E\cos(N-1)\Psi + jE\sin(N-1)\Psi$. These contributions can be added to get:

$$\begin{aligned} E_R &= E + E\cos\Psi + jE\sin\Psi + E\cos 2\Psi + jE\sin 2\Psi + \cdots \\ &= \sum_{n=0}^{N-1} E\cos n\Psi + jE\sin n\Psi \\ &= E\sum_{n=0}^{N-1} e^{jn\Psi} \end{aligned} \tag{6.37}$$

Here, N is the total number of elements in the array. A single element would have resulted in a field E, and the array is seen to modify this by the summation factor. The magnitude of summation factor is termed the *array factor* (AF):

$$\mathrm{AF} = \left|\sum_{n=0}^{N-1} e^{jn\Psi}\right| \tag{6.38}$$

The AF has a maximum value of N when $\Psi = 0$, and hence the maximum value of E_R is $E_{R\max} = NE$. Recalling that Ψ as given by Eq. (6.36) is a function of the current phase angle, α, and the angular coordinate, ϕ, it is possible to choose the current phase to make the AF show a peak in some desired direction ϕ_0. The required relationship is, from Eq. (6.36),

$$\alpha = -\frac{2\pi}{\lambda} s\cos\phi_0 \tag{6.39}$$

Combining this with Eq. (6.36) gives

$$\Psi = \frac{2\pi s}{\lambda}(\cos\phi - \cos\phi_0) \tag{6.40}$$

This can be substituted into Eq. (6.38) to give the AF as a function of ϕ relative to maximum.

Example 6.2 A dipole array has 2 elements equispaced at 0.25 wavelength. The AF is required to have a maximum along the positive axis of the array. Plot the magnitude of the AF as a function of ϕ.

Solution Given data are $N = 2$; $s = 0.25\lambda$; $\phi_0 = 0$.

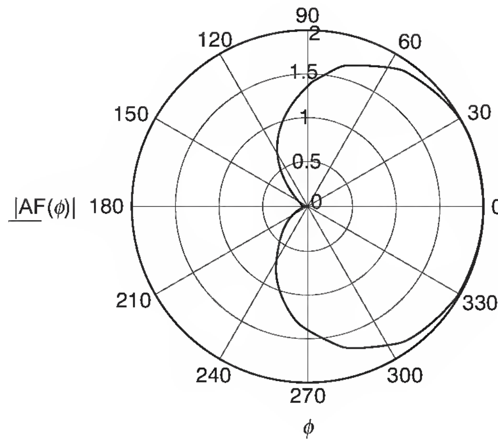
From Eq. (6.40):

$$\Psi = \frac{\pi}{2}(\cos\phi - 1)$$

The AF is

$$\begin{aligned} AF &= \left|\sum_{n=0}^{1} e^{jn\Psi}\right| \\ &= |1 + \cos\Psi + j\sin\Psi| \\ &= \sqrt{2(1 + \cos\Psi)} \end{aligned}$$

When $\phi = 0$, $\Psi = 0$, and hence the AF is 2 (as expected for two elements in the $\phi = 0$ direction). When $\phi = \pi$, $\Psi = -\pi$, and hence the AF is zero in this direction. The plot on polar graph paper is shown below.



For this particular example, the values were purposely chosen to illustrate what is termed an *end-fire array*, where the main beam is directed along the positive axis of the array. Keep in mind that a single dipole would have had a circular pattern. An example of a 5-element end fire array is given in Problem 6.32.

The current phasing can be altered to make the main lobe appear at $\phi = 90°$, giving rise to a *broadside array*. The symmetry of the dipole array means that two broadside lobes occur, one on each side of the array axis. This is illustrated in the following example.

Example 6.3 Repeat the previous example for $\phi = 90°$ and $s = 0.5\lambda$

Solution The general expression for the AF for the 2-element array is not altered and is given by

$$\text{AF} = \sqrt{2(1 + \cos\Psi)}$$

However, from Eq. (6.40), the phase angle for $s = 0.5\lambda$ and $\phi_0 = 90°$ becomes

$$\Psi = \pi\cos\phi$$

With $\phi = 0$, $\Psi = \pi$, and hence the AF is zero. Also, with $\phi = 180°$, $\Psi = -\pi$ and once again the AF is zero. With $\phi = \pm 90°$, $\Psi = 0$ and the AF is 2 in each case. The plot on polar graph paper is as shown below. An example of a 5-element broadside array is given in Problem 6.33.

As these examples show, the current phasing controls the position of the main lobe, and a continuous variation of current can be used to produce a *scanning array*. With the simple dipole array, the shape of the beam changes drastically with changes in the current phasing, and in

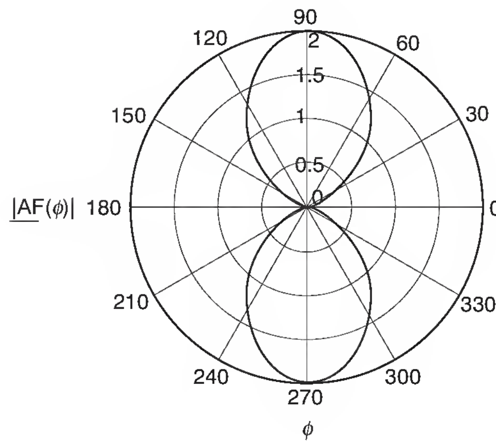




Figure 6.28 A multifeed contained-beam reflector antenna. (*Courtesy of Brain and Rudge, 1984.*)

practical scanning arrays, steps are taken to avoid this. A detailed discussion of arrays will be found in Kummer (1992).

Arrays may be used directly as antennas, and details of a nine-horn array used to provide an earth coverage beam are given in Hwang (1992). Arrays are also used as feeders for reflector antennas, and such a horn array is shown in Fig. 6.28.

6.18 Planar Antennas

A *microstrip antenna* is an antenna etched in one side of a printed circuit board, a basic *patch antenna* being as sketched in Fig. 6.29*a*. A variety of construction techniques are in use, both for the antenna pattern

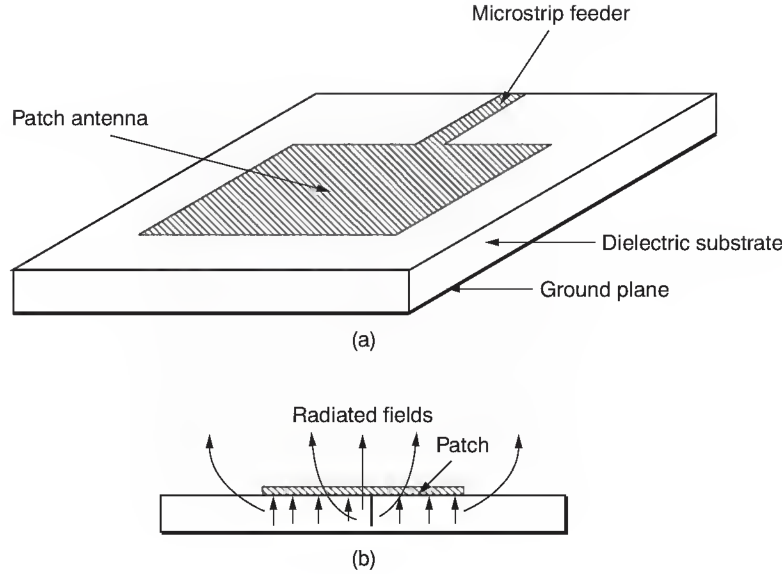


Figure 6.29 A patch antenna.

itself, and the feed arrangement, but the principles of operation can be understood from a study of the basic patch radiator. In Fig. 6.29*a* the feed is a microstrip line connecting to the patch, and the copper on the underside of the board forms a ground plane. The dielectric substrate is thin (less than about one-tenth of a wavelength) and the field under the patch is concentrated in the dielectric. At the edges of the patch the electromagnetic fields are associated with surface waves and radiated waves, the radiation taking place from the "apertures" formed in the substrate between the edges of the patch and the ground plane. The radiated fields are sketched in Fig. 6.29*b*.

Figure 6.30 shows the patch of sides a and b situated at the origin of the coordinate system of Fig. 6.3. Approximate expressions for the radiation pattern in the principal planes at $\phi = 0$ and $\phi = 90°$ are [see James et al., 1981, Eqs. (4.26a and b)]:

$$g(\theta, \phi = 90°) = \cos^2\left(\frac{\pi b}{\lambda_0}\sin\theta\right) \tag{6.41}$$

$$g(\theta, \phi = 0) = \cos^2\theta\left[\frac{\sin X}{X}\right]^2 \tag{6.42}$$

where $X = (\pi a/\lambda_0)\sin\theta$, and λ_0 is the free space wavelength. Equation (6.42) will be seen to be similar to Eq. (6.23). A plot of these functions, for a half wavelength patch is shown in Fig. 6.31. In practice the length of each side

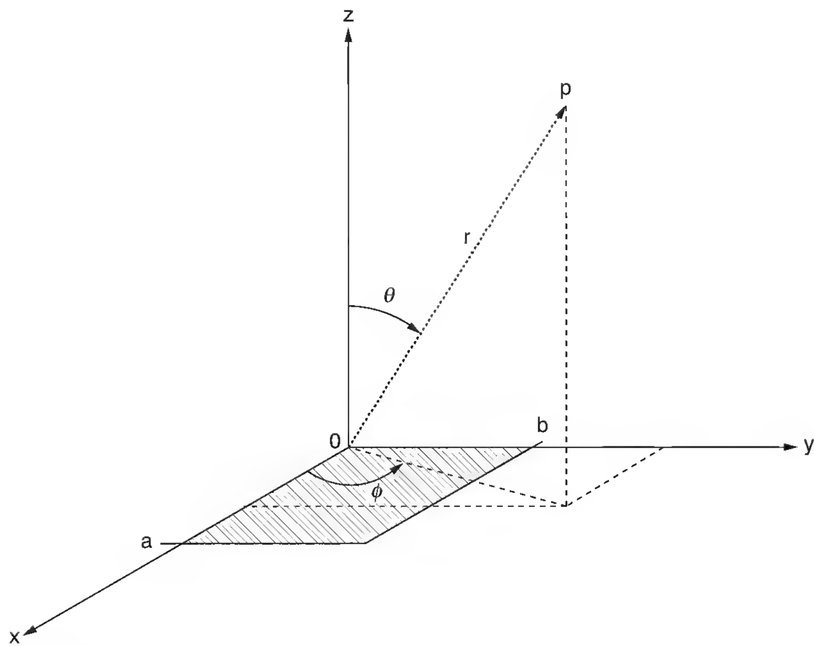


Figure 6.30 A patch antenna and its coordinate system.

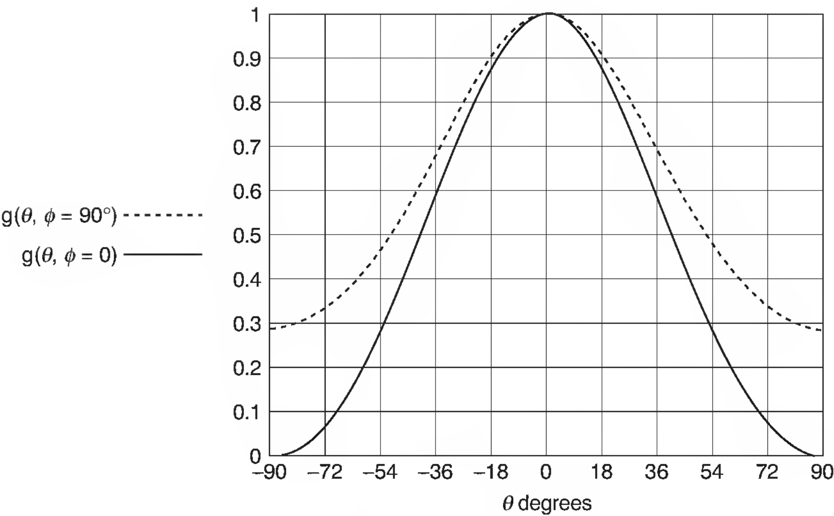


Figure 6.31 Radiation patterns for a patch antenna.

of the patch is less than half the free space wavelength because the phase velocity v_p of the wave is less than the free space value. Recall that $\lambda f = v_p$, where λ is the wavelength and f the frequency, and the phase velocity of a wave in a dielectric medium of relative permittivity ε_r is $c/\sqrt{\varepsilon_r}$, where c is the free space velocity of an electromagnetic wave. For a microstrip board of thickness 1.59 mm and a relative permittivity of 2.32, at a frequency of 10 GHz, the side length is $0.32\lambda_0$ where λ_0 is the free space wavelength [see James et al., 1981, (Table 5.3)].

Other geometries are in use, for example circular patches (disc patches), and coplanar boards and stripline boards are also used. The patch dimensions are usually half or quarter board wavelength. Figure 6.32*a* shows a disc element with a balanced coaxial feed. Figure 6.32*b* shows a *copolanar waveguide* construction. Here the board has ground planes on both sides, which are bonded together. The antenna element is etched into one of the ground planes. Figure 6.32*c* shows a *triplate* construction, Here, a *stripline*, which is etched on the inner layer of one of the boards, forms a central conductor which passes under the slot in the upper ground plane, the slot forming an aperture antenna. The two dielectrics are glued tightly together, and the ground planes are bonded together too.

The basic microstrip patch is a linear polarized antenna, but various feed arrangements are in use to convert it to a circularly polarized antenna (see, e.g., James et al., 1981).

6.19 Planar Arrays

The patch antenna is widely used in *planar arrays*. These are arrays of basic antenna elements etched on one side of a printed circuit board. A multilayer board is normally used so that associated connections and circuitry can be accommodated, as shown in Fig. 6.33. Flat panels are used, and these may be circular (as shown) or rectangular. The use of phase shifters to provide the tracking (beam scanning) is a key feature of planar arrays. The most economical method of beam forming and scanning is mechanical as described in the earlier sections. The beam is formed by a shaped reflector and azimuth and elevation motors provide the scanning. Such motors can also be used with flat panel arrays, as shown in Fig. 6.33, although the beam is formed by phasing of the elements, as described in Sec. 6.17, rather than by means of a mechanical reflector. The beam can be made to scan by introducing a progressive phase shift to the driving voltage applied to the various patch elements.

Figure 6.34 shows two basic configurations. In the active configuration, each antenna element has its own amplifier and phase shifter, while in the passive arrangement, a single amplifier drives each element

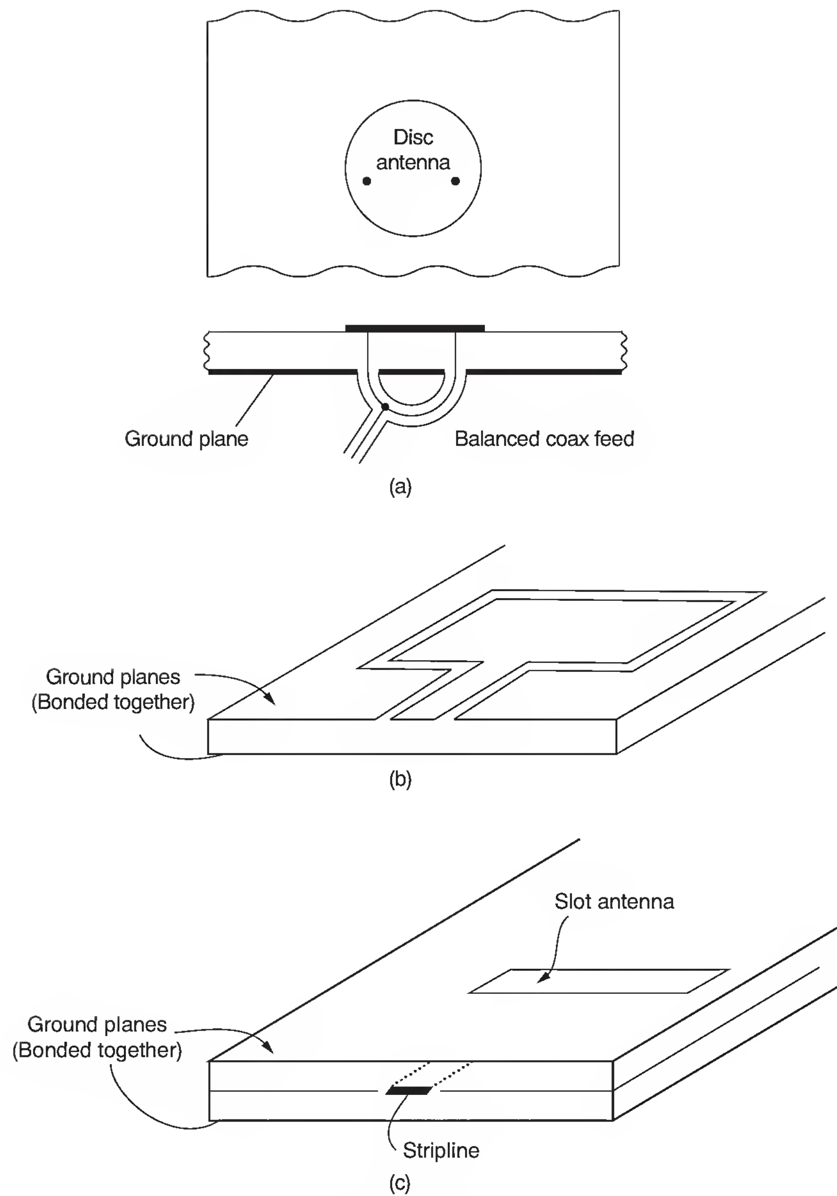


Figure 6.32 (*a*) a disc antenna with balanced coaxial feed; (*b*) a coplanar waveguide antenna; (c) a triplate slot antenna.

through the individual phase shifters. The active arrangement has the advantage that the failure of one amplifier causes degradation but not complete loss of signal, but this has to be offset against the greater cost and complexity incurred.

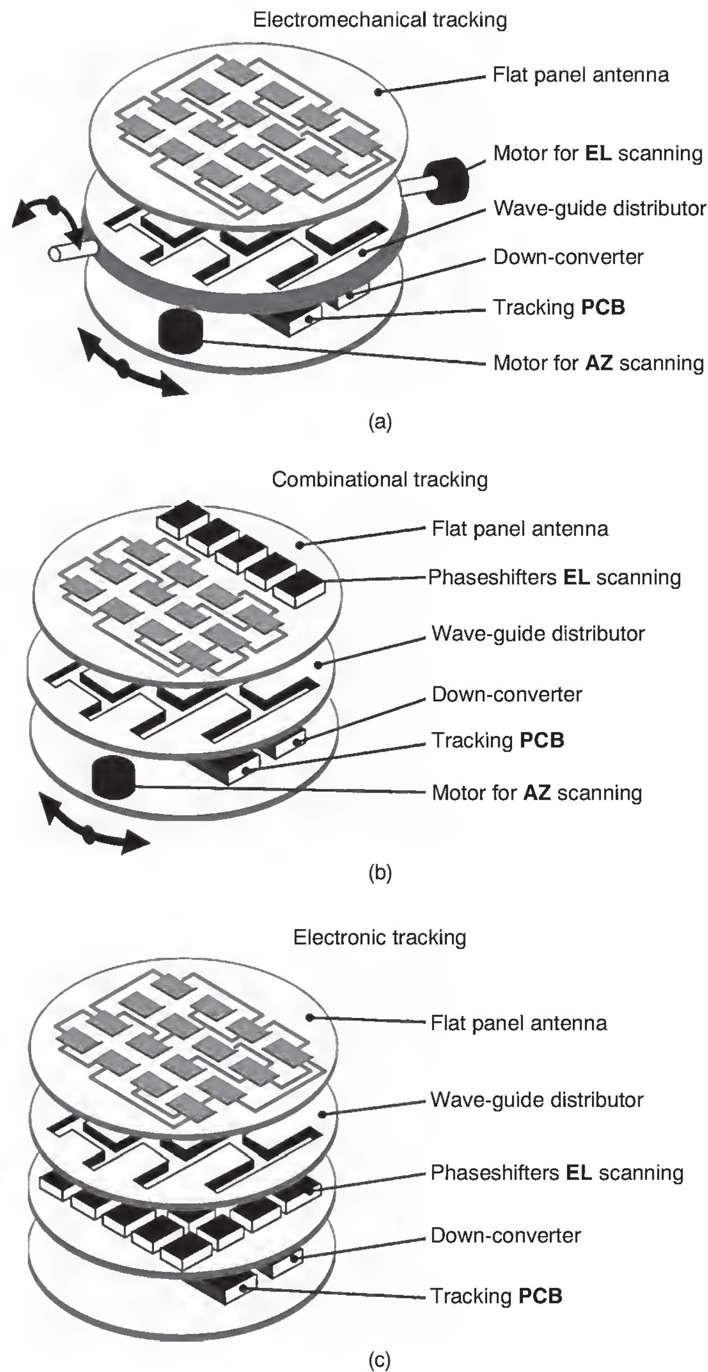


Figure 6.33 Assembly details of a planar array. (*a*) electromechanical tracking, (*b*) combined electromechanical and electronic tracking, (*c*) electronic tracking. (*Courtesy of Michael Parnes. Source: http://www.ascor.eltech.ru/ascor15.htm*)

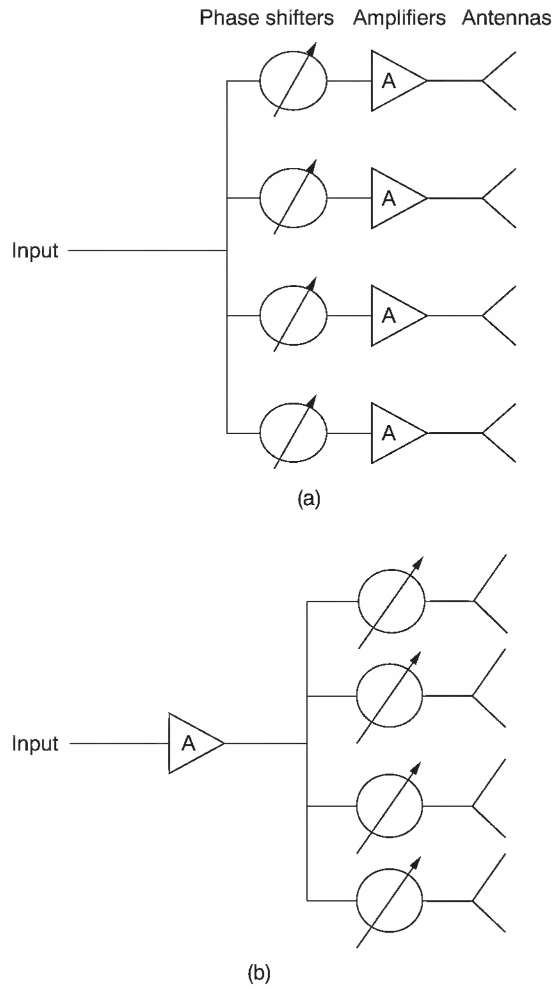


Figure 6.34 (*a*) Active, and (*b*) passive array configurations.

Electronic scanning can be achieved in one of several ways. The phase shift coefficient of a transmission line is given by $\beta = 2\pi/\lambda$, where λ is the wavelength of the signal passing along the line. A section of transmission line of length l introduces a phase lag (the output lags the input) of amount

$$\varphi = \beta l \tag{6.43}$$

Phase shift can be achieved by changing either l or β. The concept of changing the length l is illustrated in Fig. 6.35. Making the shorter

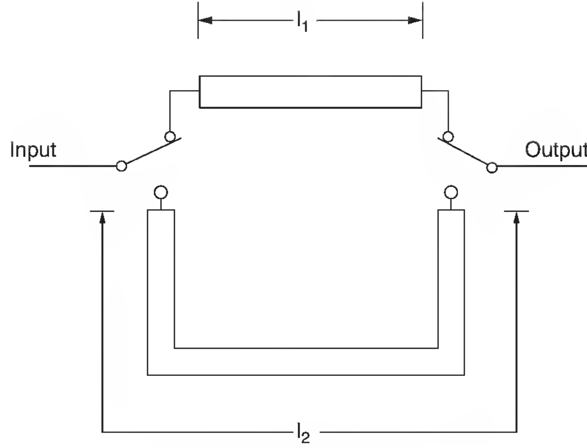


Figure 6.35 A transmission line phase shifter.

length the reference line, the phase shift obtained in switching from one to the other is

$$\Delta\varphi = \beta(l_1 - l_2) \tag{6.44}$$

It will be seen that the switched line phase shifter requires a *double pole single throw* (DPST) switch at each end. Several types of switches have been utilized in practical designs, including PIN diodes, *field effect transistors* (FETs) and *micro-electro-mechanical* (MEM) switches. In a PIN diode, the *p*-type semiconductor region is separated from the *n*-type region by an intrinsic region (hence the name PIN). At frequencies below about 100 MHz, the diode behaves as a normal rectifying diode. Above this frequency, the stored charge in the intrinsic region prevents rectification from occurring and the diode conducts in both directions. The diode resistance is inversely related to the stored charge, which in turn is controlled by a steady bias voltage. With full forward bias the diode appears as a short circuit, and with full reverse bias the diode ceases to conduct. In effect the diode behaves as a switch.

In practice PIN diode switches are usually wire-bonded into the phase changer, this being referred to as a *microwave integrated circuit* (MIC). The wire bond introduces a parasitic inductance which sets an upper frequency limit, although they have been used at frequencies beyond 18 GHz. Two diodes are required for each DPST switch.

Metal semiconductor field effect transistors (MESFETs) are also widely used as microwave switches. In the MESFET, the charge in the channel between the drain and source electrodes is controlled by the bias voltage applied to the gate electrode. The channel can be switched between a

highly conducting (ON) state and a highly resistive (OFF) state. MESFETs utilize gallium arsenide (GaAs) substrates, and can be constructed along with the line elements as an integrated circuit, forming what is known as a *monolithic microwave integrated circuit* (MMIC). (MMICs may also contain other active circuits such as amplifiers and oscillators). Figure 6.36 shows four MESFETS integrated into a switched line phase shifter.

The MEM switch is a small ON/OFF type switch that is actuated by electrostatic forces. In one form, a cantilever gold beam is suspended over a control electrode, these two elements forming an air-spaced capacitor. The dimensions of the beam are typically in the range of a few hundred microns (1 micron, abbreviated 1 μm is 10^{-6} meters) with an air gap of a few microns.

The RF input is connected to one end of the beam, which makes contact with an output electrode when the beam is pulled down. The pull down action occurs as a result of the electrostatic force arising when a direct voltage is applied between the control electrode and the beam. The voltage is in the order of 75 V, but little current is drawn. The power required to activate the switch depends on the number of cycles per second and the capacitance. In one example (see Reid, 2005), a voltage of 75 V, capacitance of 0.5 pF and a switching frequency of 10 kHz resulted in a power requirement of 14 μW.

A MEM switch can also be constructed where the beam, fixed at both ends, forms an air bridge across the control electrode (Brown, 1998). The top surface of the control electrode has a thin dielectric coating. The electrostatic force deflects the beam causing it to clamp down on the dielectric coating. The capacitance formed by the beam, dielectric coating,

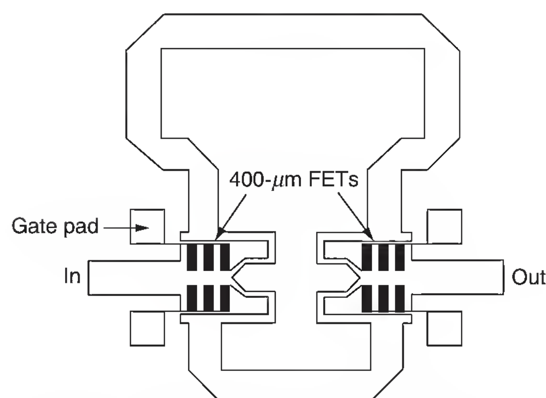


Figure 6.36 A MESFET switched line phase shifter. (*From* http://parts.jpl.nasa.gov/mmic/3-IX.PDF)

and control electrode provides the RF coupling between input and output. Thus, this is basically a contactless switch.

As mentioned earlier, it is also possible to alter the phase shift by altering the propagation coefficient. By definition, a sinusoidal electromagnetic wave experiences a phase change of 2π rad over distance of one wavelength λ, and therefore the phase change coefficient can be written simply as

$$\beta = \frac{2\pi}{\lambda} \tag{6.45}$$

This will be in radians per meter with λ in meters. As noted earlier, the connection between wavelength λ, frequency f, and phase velocity v_p is $\lambda f = v_p$. It is also known that the phase velocity on a transmission line having a dielectric of relative permittivity ε_r is $v_p = c/\sqrt{\varepsilon_r}$, where c is the free space velocity of light. Substituting these relationships in Eq. (6.43) gives:

$$\varphi = \frac{2\pi f\sqrt{\varepsilon_r}}{c}l \tag{6.46}$$

This shows that, for a fixed length of line, a phase change can be obtained by changing the frequency f or by changing the relative permittivity (dielectric constant) ε_r. In one scheme (Nishio et al., 2004) a method is given for phasing a base station antenna array by means of frequency change. The modulated subcarrier is fed in parallel to a number of heterodyne frequency mixers. A common *local oscillator* (LO) signal is fed to each mixer to change the subcarrier up to the assigned carrier frequency, the output from each mixer feeding its own element in the antenna array. The phase change is introduced into the LO circuit by having a different, fixed length of line in each branch of the LO feed to the mixers. Thus the output from each mixer will have its own fixed phase angle, determined by the phase shift in the oscillator branch.

Phase change can also be effected by changing the relative permittivity of a delay line. Efforts in this direction have concentrated on using *ferroelectric* material as a dielectric substrate for the delay line. Whereas the dielectric constant of a printed circuit board may range from about 2 to 10, ferroelectrics have dielectric constants measured in terms of several hundreds. The ferroelectric dielectric constant can be changed by application of an electric field, which may be in the order of 2000 kV/m. Thus to keep the applied voltage to reasonable levels, a thin dielectric is needed. For example, for a dielectric thickness of 0.15 mm and an electric field of 2000 kV/m the applied voltage

is $2000 \times 10^3 \times 0.15 \times 10^{-3} = 300$ V. The characteristic impedance of a microstrip line is given by $Z_0/\sqrt{\varepsilon_r}$, where Z_0 is the impedance of the same line with an air dielectric. The characteristic impedance increases as a function of h/W, where W is the width of the line and h the thickness of the dielectric. Thus thinner dielectrics lead to lower characteristic impedance, and this combined with the high dielectric constant means that the line width W has to be narrow. Values given in De Flaviis et al. (1997) for the ferroelectric material barium modified strontium titanium oxide ($Ba_{1-x} Sr_x TiO_3$) show a dielectric constant in the region of 600, dielectric thickness between 0.1 and 0.15 mm, and line width of 50 μm, for a characteristic impedance of 50 Ω. The bias voltage is 250 V.

The ferroelectric dielectric is used in a number of different ways. In the paper by De Flaviis et al., the material was used simply as the dielectric for a microstrip delay line. It has also been used as a lens to produce scanning by deflecting an antenna beam (see Ferroelectric Lens Phased Array at http://radar-www.nrl.navy.mil/Areas/Ferro). The lens is shown in Fig. 6.37. The ferroelectric dielectric in each column is biased to provide a progressive phase shift so that an incident plane wave normal to the edges of the dielectric columns will emerge from the opposite edges in a direction determined by the phasing in the lens. A single lens produces one dimensional scanning. Ferroelectrics are also employed in reflectarrays described next.

6.20 Reflectarrays

A *reflectarray*, as the name suggests is an array of antenna elements that acts as a reflector. A reflector array incorporates a planar array as a reflector, as shown in Fig. 6.38. The planar array basically replaces the parabolic reflector shown in Fig. 6.13. Reflected waves from each of the elements in the array can be phased to produce beam scanning; Fig. 6.38*b* shows the construction. The reflected wave is actually a combination of reflections from the antenna elements and the substrate. Figure 6.38*c* shows the polar diagram for a 784 element array. Further details of this array will be found in Pozar (2004).

A number of methods of producing beam scanning have been proposed. Fig. 6.39 shows a 2832 element, 19-GHz reflectarray which employs ferroelectric phase shifters for the elements. Further details will be found at www.ctsystemes.com/zeland/publi/TM-2000-210063.pdf. Varactor diodes have also been used to provide a phase shift that is controlled by an applied bias. A varactor is in effect a voltage controlled capacitor, and changes in the capacitance introduce a corresponding phase shift. Figure 6.40 shows one arrangement for a five-element array.

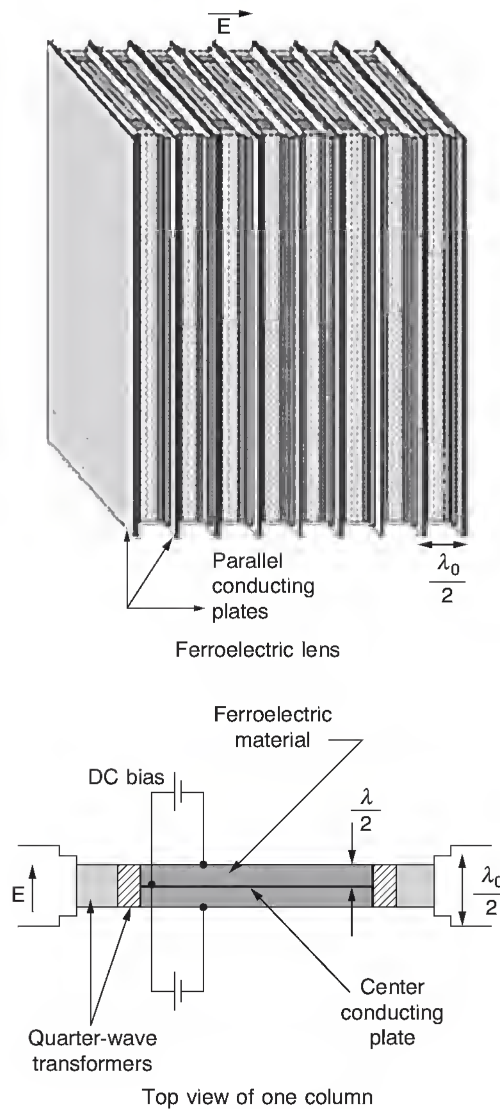
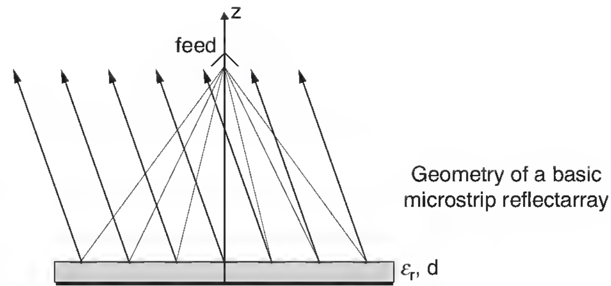


Figure 6.37 A ferroelectric lens. (*Courtesy of U. S. Naval Research Laboratory Radar Division, Washington, DC. Source: http://radar-www.nrl.navy.mil/Areas/Ferro*)

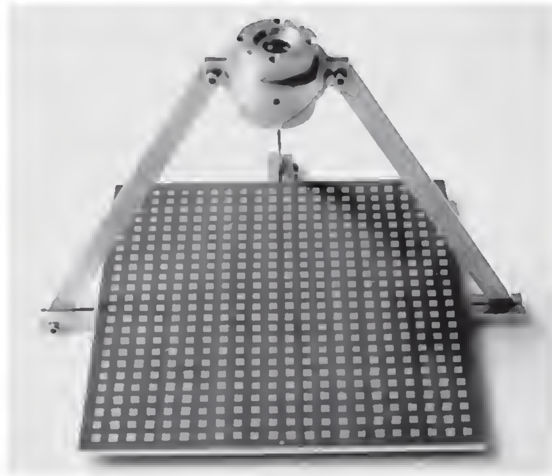
Further details will be found at www.ansoft.com/news/articles/ICEAA2001.pdf.

6.21 Array Switching

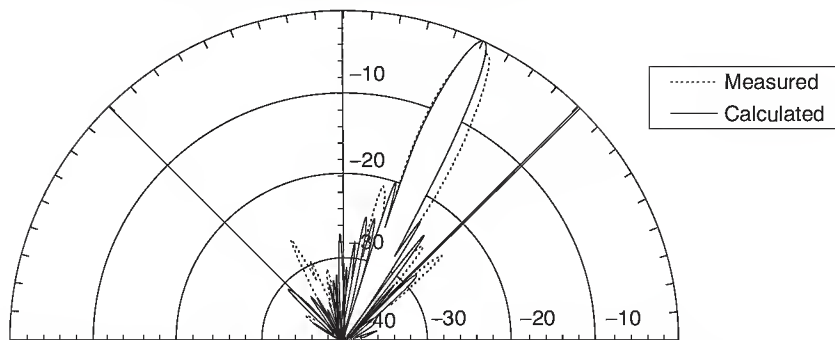
Switching of the phasing elements in the antenna arrays is usually carried out digitally. (There are analog phase shifters, some employing ferrite materials, which offer continuously variable phase change, but



Example of a 28 GHz microstrip reflectarray
(using variable size patches)



Patterns of 28 GHz reflectarray



28 GHz, 6" square aperture, 784 elements (variable size patches), 25 degree scan angle, corrugated conical horn feed, G = 31 dB, 51% aperture efficiency

Figure 6.38 An 784 element reflectarray. (*Courtesy of D.M. Pozar. Source: www.ecs.umass.edu/ece/pozar/jina.ppt*)

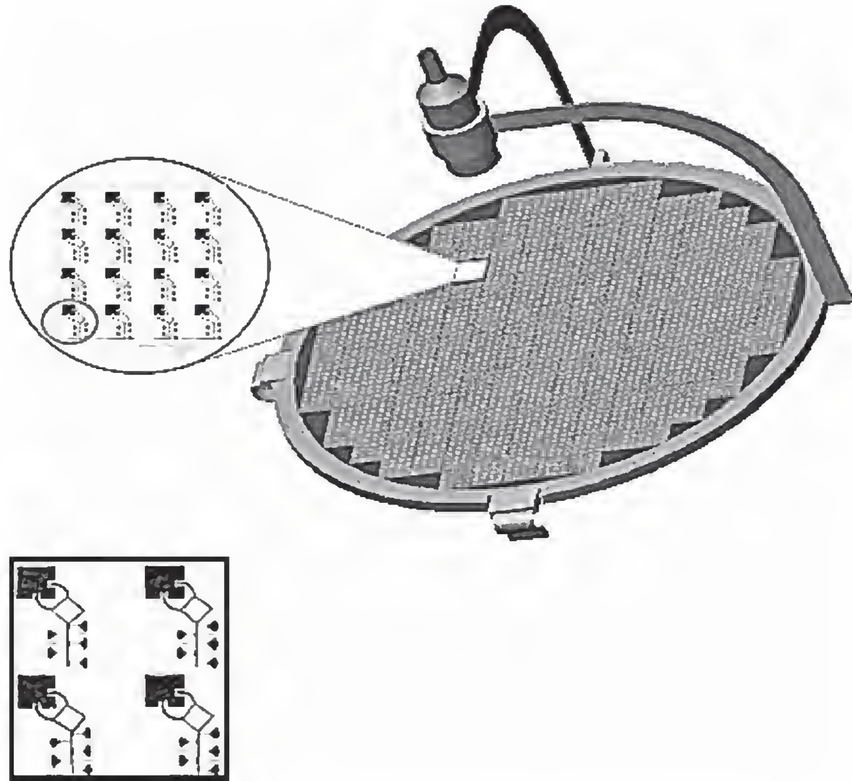


Figure 6.39 A 2832 element 19 GHz ferroelectric reflectarray concept. The callout shows a 16 element subarray patterned on a 3.1 × 3.1 cm, 0.25 mm thick MgO substrate. The array diameter is 48.5 cm. The unit cell area is 0.604 cm^2 and the estimated boresight gain is 39 dB. (*Courtesy of Robert R. Romanofsky, NASA Glenn Research Center, Cleveland Ohio. Source: gltrs.grc.nasa.gov/reports/2000/TM-2000-210063.pdf*)

these are not considered here). The digitally switched delay line type offers faster switching speed which is an important consideration where beam scanning is employed. The phase shift increments are determined by successive division by 2 of 360°. These would follow the pattern 180°, 90°, 45°, 22.5°, 12.25°, and so on. Thus a 4-bit phase shifter would have $2^4 = 16$ states, providing increments of 180°, 90°, 45°, 22.5°. This is another limitation of digitally switched phase shifters, the resolution that can be achieved. One manufacturer (KDI Triangle Corp. states that 5.63° is the practical limit for digital, compared to 0.088° for analog types). The arrangement for a 4-bit phase shifter is shown in Fig. 6.41*a*. It is seen that four control lines are required, one for each logical bit. The *most significant bit* (MSB) switches in the 180° phase shift, and the *least significant bit* (LSB) the 22.5°, and if all four lines

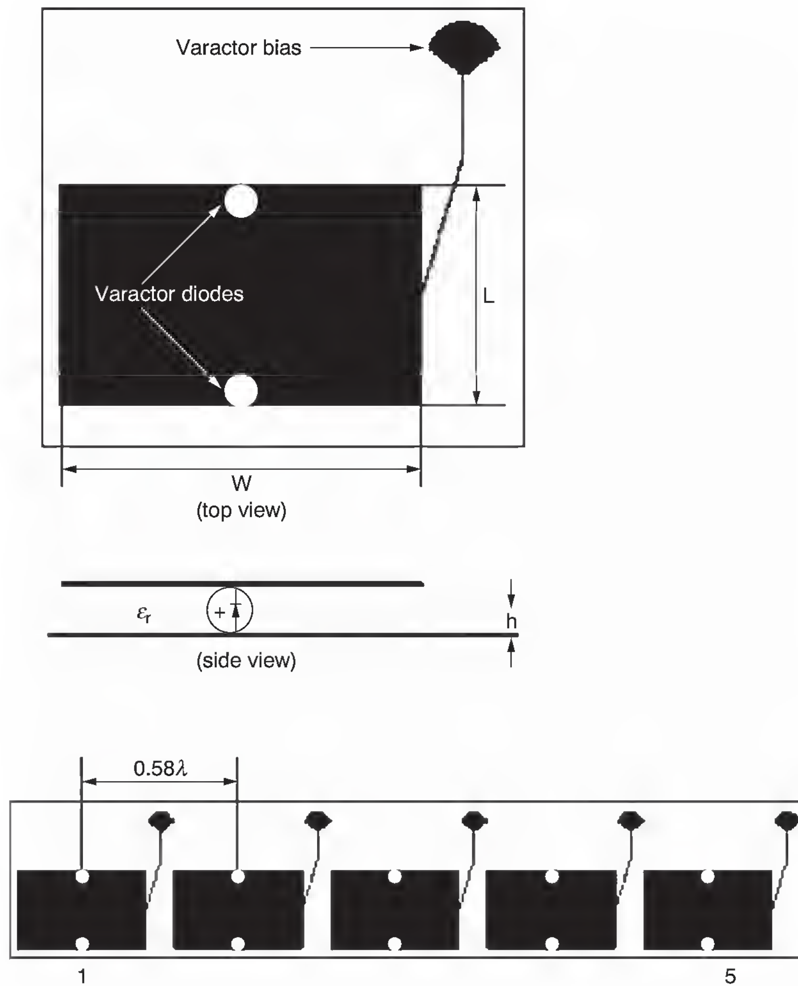
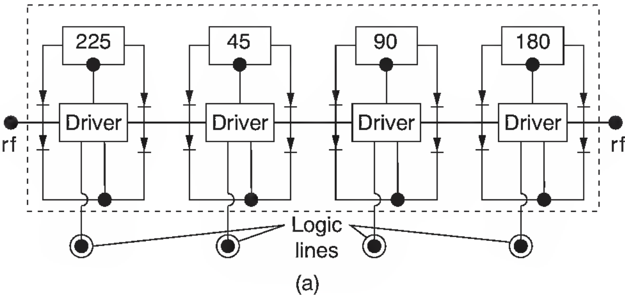


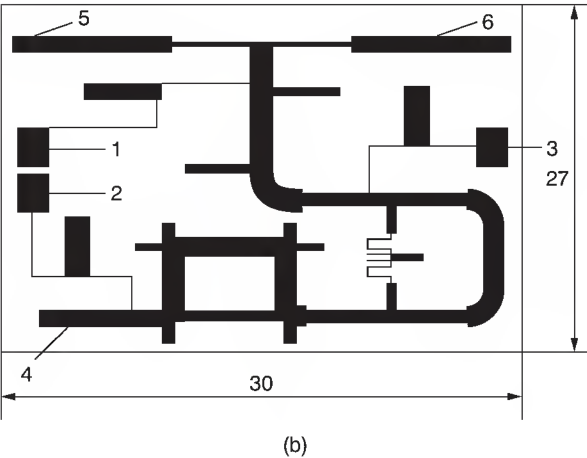
Figure 6.40 Phase shift using varactor diodes. (*Courtesy of L. Boccia. Source: www.ansoft.com/news/articles/ICEAA2001.pdf*)

are activated the phase shift is the sum of the four delays, or 337.5°. With all four delay lines out of circuit the phase shift is back to zero, or 360°.

Care must be taken how "bits" are interpreted. A "three bit" line would have a MSB of 180° and a LSB of 45° and would require three logic control lines. However in the 3-bit phase shifter shown in Fig. 6.41*b* positive and negative control voltages can be applied to the control lines, giving rise to the phase shifts shown in the table of Fig. 6.41*c*.



(a)



(b)

Nominal phase state	LEAD connections		
	cont. pl. 1	cont. pl. 2	cont. pl. 3
45	+	–	+
90	–	+	+
135	+	+	+
180	–	–	–
225	+	–	–
270	–	+	–
315	+	+	–

Terminology: "+" - control current "–" - reverse voltage.

(c)

Figure 6.41 Digital switching of phased arrays: (*a*) 4-bit (*Courtesy of T. J. Braviak. Source: www.kditriangle.com*), (*b*) Ku-band p-i-n diode (3-bit) phase shifter, (*c*) the switching logic for (*b*). (*Courtesy of Michael Parnes. Source: http://www.ascor.eltech.ru/ascor15.htm*)

6.22 Problems and Exercises

6.1. The power output from a transmitter amplifier is 600 W. The feeder losses amount to 1 dB, and the voltage reflection coefficient at the antenna is 0.01. Calculate the radiated power.

6.2. Explain what is meant by the *reciprocity theorem* as applied to antennas. A voltage of 100 V applied at the terminals of a transmitting dipole antenna results in an induced current of 3 mA in a receiving dipole antenna. Calculate the current induced in the first antenna when a voltage of 350 V is applied to the terminals of the second antenna.

6.3. The position of a point in the coordinate system of Sec. 6.3 is given generally as $r(\theta, \phi)$. Determine the x, y, and z coordinates of a point 3(30°, 20°).

6.4. What are the main characteristics of a radiated wave in the far-field region? The components of a wave in the far field region are $E_\theta = 3$ mV/m, $E_\phi = 4$ mV/m. Calculate the magnitude of the total electric field. Calculate also the magnitude of the magnetic field.

6.5. The **k** vector for the wave specified in Prob. 6.4 is directed along the $+x$ axis. Determine the direction of the resultant electric field in the yz plane.

6.6. The **k** vector for the wave specified in Prob. 6.4 is directed along the $+z$ axis. Is there sufficient information given to determine the direction of the resultant electric field in the xy plane? Give reason for your answer.

6.7. The rms value of the electric field of a wave in the far-field region is 3 μV/m. Calculate the power flux density.

6.8. Explain what is meant by the *isotropic power gain* of an antenna. The gain of a reflector antenna relative to a $1/2\lambda$-dipole feed is 49 dB. What is the isotropic gain of the antenna?

6.9. The directivity of an antenna is 52 dB, and the antenna efficiency is 0.95. What is the power gain of the antenna?

6.10. The radiation pattern of an antenna is given by $g(\theta, \phi) = |\sin\theta \sin\phi|$. Plot the resulting patterns for (*a*) the xz plane and (*b*) the yz plane.

6.11. For the antenna in Prob. 6.10, determine the half-power beamwidths, and hence determine the directivity.

6.12. Explain what is meant by the *effective aperture* of an antenna. A paraboloidal reflector antenna has a diameter of 3 m and an illumination efficiency of 70 percent. Determine (*a*) its effective aperture and (*b*) its gain at a frequency of 4 GHz.

6.13. What is the effective aperture of an isotropic antenna operating at a wavelength of 1 cm?

6.14. Determine the half-power beamwidth of a half-wave dipole.

6.15. A uniformly illuminated rectangular aperture has dimensions $a = 4\lambda$, $b = 3\lambda$. Plot the radiation patterns in the principal planes.

6.16. Determine the half-power beamwidths in the principal planes for the uniformly illuminated aperture of Prob. 6.15. Hence determine the gain. State any assumptions made.

6.17. Explain why the smooth-walled conical horn radiates copolar and cross-polar field components. Why is it desirable to reduce the cross-polar field as far as practical, and state what steps can be taken to achieve this.

6.18. When the rectangular aperture shown in Fig. 6.9 is fed from a waveguide operating in the TE_{10} mode, the far-field components (normalized to unity) are given by

$$E_\theta(\theta, \phi) = -\frac{\pi}{2}\sin\phi\frac{\cos X}{X^2 - \left(\frac{\pi}{2}\right)^2}\frac{\sin Y}{Y}$$

$$E_\phi(\theta, \phi) = E_\theta(\theta, \phi)\cos\theta\cot\phi$$

where X and Y are given by Eqs. (6.19) and (6.20). The aperture dimensions are $a = 3\lambda$, $b = 2\lambda$. Plot the radiation patterns in the principal planes.

6.19. Determine the half-power beamwidths in the principal planes for the aperture specified in Prob. 6.18, and hence determine the directivity.

6.20. A pyramidal horn antenna has dimensions $a = 4\lambda$, $b = 2.5\lambda$, and an illumination efficiency of 70 percent. Determine the gain.

6.21. What are the main characteristics of a parabolic reflector that make it highly suitable for use as an antenna reflector?

6.22. Explain what is meant by the *space attenuation function* in connection with the paraboloidal reflector antenna.

6.23. Figure 6.17*b* can be referred to *xy* rectangular coordinates with *A* at the origin and the *x* axis directed from *A* to *S*. The equation of the parabola is then $y^2 = 4fx$. Given that $y_{max} = \pm 2.5$ m at $x_{max} = 0.9$ m, plot the space attenuation function.

6.24. What is the *f*/*D* ratio for the antenna of Prob. 6.23? Sketch the position of the focal point in relation to the reflector.

6.25. Determine the depth of the reflector specified in Prob. 6.23.

6.26. A 3-m paraboloidal dish has a depth of 1 m. Determine the focal length.

6.27. A 5-m paraboloidal reflector works with an illumination efficiency of 65 percent. Determine its effective aperture and gain at a frequency of 6 GHz.

6.28. Determine the half-power beamwidth for the reflector antenna of Prob. 6.27. What is the beamwidth between the first nulls?

6.29. Describe briefly the *offset feed* used with paraboloidal reflector antennas, stating its main advantages and disadvantages.

6.30. Explain why double-reflector antennas are often used with large earth stations.

6.31. Describe briefly the main advantages to be gained in using an antenna array.

6.32. A basic dipole array consists of five equispaced dipole elements configured as shown in Fig. 6.26. The spacing between elements is 0.3λ. Determine the current phasing needed to produce an end-fire pattern. Provide a polar plot of the AF.

6.33. What current phasing would be required for the array in Prob. 6.32 to produce a broadside pattern?

6.34. A four-element dipole array, configured as shown in Fig. 6.26, is required to produce maximum radiation in a direction $\phi_0 = 15°$. The elements are spaced by 0.2λ. Determine the current phasing required, and provide a polar plot of the AF.

6.35. A rectangular patch antenna element has sides $a = 9$ mm, $b = 6$ mm. The operating frequency is 10 GHz. Plot the radiation patterns for the $\phi = 0$ and $\phi = 90°$ planes.

6.36. For microstrip line, where the thickness t of the line is negligible compared to the dielectric thickness h, and the line width $W \geq h$ the effective dielectric constant is given by

$$\varepsilon_e \cong \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2\sqrt{1 + 12\dfrac{h}{W}}}$$

ε_r is the dielectric constant of the dielectric material. The characteristic impedance is given by

$$Z_0 = \frac{120\pi}{\sqrt{\varepsilon_e}}\left[\frac{W}{h} + 1.393 + 0.667\ln\left(\frac{W}{h} + 1.444\right)\right]^{-1}$$

(see Chang, 1989). Calculate the characteristic impedance for a microstrip line of width 0.7 mm, on an alumina dielectric of thickness 0.7 mm. The dielectric constant is 9.7.

6.37. For the microstripline of Prob. 6.36, calculate (*a*) the line wavelength (*b*) the phase shift coefficient in rad/m, and in degrees/cm. The frequency of operation is 10 GHz.

6.38. The dielectric constant of polyguide dielectric is 2.32. Calculate the characteristic impedance and phase shift coefficient for a microstrip line of width 2.45 mm, and dielectric thickness 1.58 mm.

6.39. The effective dielectric constant for a microstripline is 1.91. Design a switched-line phase shifter (see Fig. 6.35) to produce a phase shift of 22.5° at a frequency of 12 GHz. Show how switching might be achieved using PIN diodes.

6.40. Calculate the power required to drive a MEM switch, which has to operate at a frequency of 8 kHz. The switch capacitance is 0.5 pF, and the drive voltage needed for switching is 75 V. (*Hint*: The energy stored in a capacitor is 1/2 CV^2 and power is J/s.)

References

Balanis, C. 1982. *Antenna Theory Analysis and Design.* Harper & Row, New York.

Brain, D. J., and A. W. Rudge. 1984. "Electronics and Power." *J. of the IEEE*, Vol. 30, No. 1, January, pp. 51–56.

Brown, R. E. 1998. "RF-MEMS Switches for Reconfigurable Integrated Circuits." *IEEE Trans. on Microwave Theory and Techniques*, Vol. 46, No. 11, November, pp. 1868–1880.

Chang, K. (ed.). 1989. *Handbook of Microwave and Optical Components,* Vol. 1. Wiley, New York.

De Flaviis, F., N. G. Alexopoulos, and, O. M. Stafsudd. 1997. "Planar Microwave Integrated Phase Shifter Design with High Purity Ferroelectric Material." *IEEE Trans. on Microw. Theory and Tech.*, Vol. 45, No. 6, June, pp. 963–969 (see also www.ece.uci.edu/rfmems/publications/papers-pdf/J005.PDF).

Glazier, E. V. D., and H. R. L. Lamont. 1958. *The Services Textbook of Radio*, Vol. 5: *Transmission and Propagation.* Her Majesty's Stationery Office, London.

Hwang, Y. 1992. "Satellite Antennas." *Proc. IEEE*, Vol. 80, No. 1, January, pp. 183–193.

James, J. R., P. S. Hall, and C. Wood. 1981. *Microstrip Antenna Theory and Design.* Peter Peregrinus, UK.

Kummer, W. H. 1992. "Basic Array Theory." *Proc. IEEE*, Vol. 80, No. 1, January, pp. 127–140.

Miya, K. (ed.). 1981. *Satellite Communications Technology.* KDD Engineering and Consulting, Japan.

Olver, A. D. 1992. "Corrugated Horns." *Electron. Commun. Eng. J.*, Vol. 4, No. 10, February, pp. 4–10.

Pozar, D. M.2004. "Microstrip Reflectarrays Myths and Realities." JINA Conference (see www.ecs.umass.edu/ece/pozar/jina.ppt).

Reid, J. R. 2005. "Microelectromechanical Phase Shifters for Lightweight Antennas," at www.afrlhorizons.com.

Rudge, A. W., K. Milne, A. D. Olver, and P. Knight (eds.). 1982. *The Handbook of Antenna Design*, Vol. 1. Peter Peregrinus, UK.

Rusch, W. V. T. 1992. "The Current State of the Reflector Antenna Art: Entering the 1990s." *Proc. IEEE*, Vol. 80, No. 1, January, pp. 113–126.

Nishio, T., X. Hao, W. Yuanxun, and T. Itoh, 2004. "A Frequency Controlled Active Phased Array." *IEEE Microw. and Components Lett.*, Vol. 14, No. 3 March. pp. 115–117.

Vectors. 1993. Hughes in-house magazine, Vol. XXXV, No. 3.

Chapter

7

The Space Segment

7.1 Introduction

A satellite communications system can be broadly divided into two segments—a ground segment and a space segment. The space segment will obviously include the satellites, but it also includes the ground facilities needed to keep the satellites operational, these being referred to as the *tracking, telemetry, and command* (TT&C) facilities. In many networks it is common practice to employ a ground station solely for the purpose of TT&C.

The equipment carried aboard the satellite also can be classified according to function. The *payload* refers to the equipment used to provide the service for which the satellite has been launched. The *bus* refers not only to the vehicle which carries the payload but also to the various subsystems which provide the power, attitude control, orbital control, thermal control, and command and telemetry functions required to service the payload.

In a communications satellite, the equipment which provides the connecting link between the satellite's transmit and receive antennas is referred to as the *transponder*. The transponder forms one of the main sections of the payload, the other being the antenna subsystems.

In this chapter the main characteristics of certain bus systems and payloads are described.

7.2 The Power Supply

The primary electrical power for operating the electronic equipment is obtained from solar cells. Individual cells can generate only small amounts of power, and therefore, arrays of cells in series-parallel connection are required. Figure 7.1 shows the solar cell panels for the HS 376 satellite

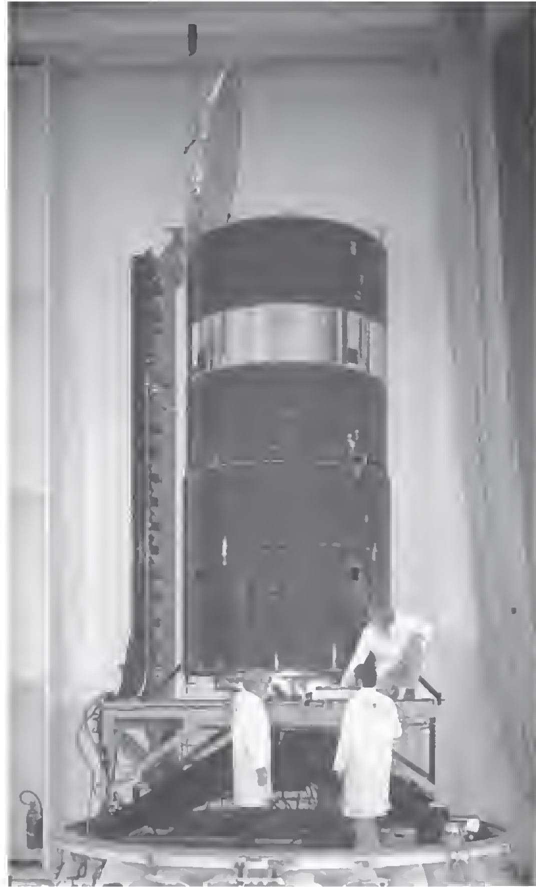


Figure 7.1 The HS 376 satellite. (*Courtesy of Hughes Aircraft Company Space and Communications Group.*)

manufactured by Hughes Space and Communications Company. The spacecraft is 216 cm in diameter and 660 cm long when fully deployed in orbit. During the launch sequence, the outer cylinder is telescoped over the inner one, to reduce the overall length. Only the outer panel generates electrical power during this phase. In geostationary orbit the telescoped panel is fully extended so that both are exposed to sunlight. At the beginning of life, the panels produce 940 W dc power, which may drop to 760 W at the end of 10 years. During eclipse, power is provided by two nickel-cadmium (Ni-Cd) long-life batteries, which will deliver 830 W. At the end of life, battery recharge time is less than 16 h.

The HS 376 spacecraft is a spin-stabilized spacecraft (the gyroscopic effect of the spin is used for mechanical orientational stability, as

described in Sec. 7.3). Thus the arrays are only partially in sunshine at any given time, which places a limitation on power.

Higher powers can be achieved with solar panels arranged in the form of rectangular *solar sails*. Solar sails must be folded during the launch phase and extended when in geostationary orbit. Figure 7.2 shows the HS 601 satellite manufactured by Hughes Space and Communications Company. As shown, the solar sails are folded up on each side, and when fully extended, they stretch to 67 ft (20.42 m) from tip to tip. The full complement of solar cells is exposed to the sunlight, and the sails are arranged to rotate to track the sun, so they are capable of greater power output than cylindrical arrays having a comparable number of cells. The HS 601 can be designed to provide dc power from 2 to 6 kW. In comparing the power

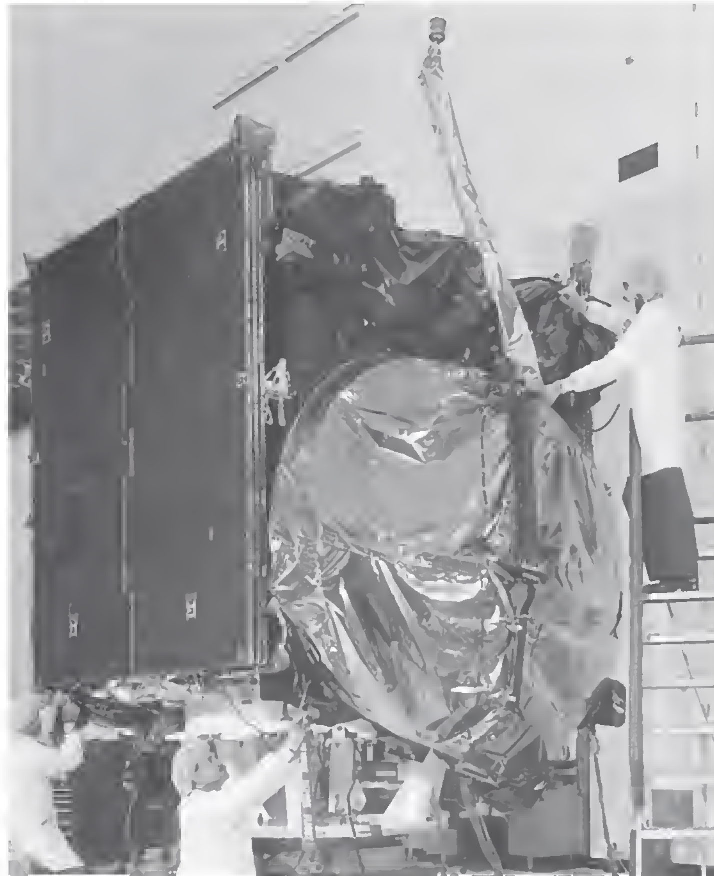


Figure 7.2 Aussat B1 (renamed Optus B), Hughes first HS 601 communications satellite is prepared for environmental testing. (*Courtesy of Hughes Aircraft Company Space and Communications Group.*)

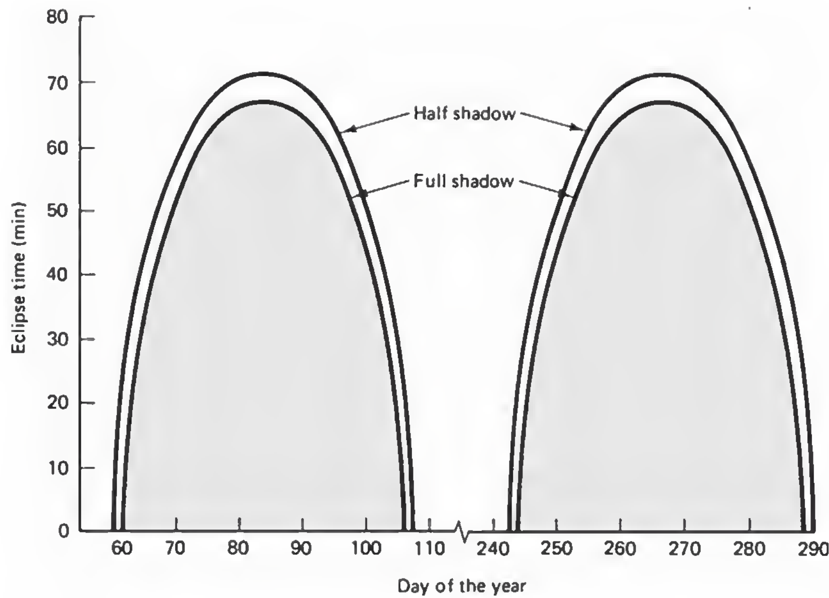


Figure 7.3 Satellite eclipse time as a function of the current day of the year. (*Courtesy of Spilker, 1977. Reprinted by permission of Prentice-Hall, Englewood Cliffs, NJ.*)

capacity of cylindrical and solar-sail satellites, the cross-over point is estimated to be about 2 kW, where the solar-sail type is more economical than the cylindrical type (Hyndman, 1991).

As discussed in Sec. 3.6, the earth will eclipse a geostationary satellite twice a year, during the spring and autumnal equinoxes. Daily eclipses start approximately 23 days before and end approximately 23 days after the equinox for both the spring and autumnal equinoxes and can last up to 72 min at the actual equinox days. Figure 7.3 shows the graph relating eclipse period to the day of year. In order to maintain service during an eclipse, storage batteries must be provided. Ni-Cd batteries continue to be used, as shown in the Hughes HS 376 satellite, but developments in nickel-hydrogen (Ni-H_2) batteries offer significant improvement in power-weight ratio. Ni-H_2 batteries are used in the Hughes HS 601 (e.g., the SATMEX-5 and Anik-F2 satellites, see Secs. 7.9 and 7.10) and were introduced into the Intelsat series with INTELSAT VI (Pilcher, 1982) and INTELSAT VII (Lilly, 1990) satellites.

7.3 Attitude Control

The *attitude* of a satellite refers to its orientation in space. Much of the equipment carried aboard a satellite is there for the purpose of controlling its attitude. Attitude control is necessary, for example, to ensure that

directional antennas point in the proper directions. In the case of earth environmental satellites, the earth-sensing instruments must cover the required regions of the earth, which also requires attitude control. A number of forces, referred to as *disturbance torques*, can alter the attitude, some examples being the gravitational fields of the earth and the moon, solar radiation, and meteorite impacts. Attitude control must not be confused with station keeping, which is the term used for maintaining a satellite in its correct orbital position, although the two are closely related.

To exercise attitude control, there must be available some measure of a satellite's orientation in space and of any tendency for this to shift. In one method, infrared sensors, referred to as *horizon detectors*, are used to detect the rim of the earth against the background of space. With the use of four such sensors, one for each quadrant, the center of the earth can be readily established as a reference point. Any shift in orientation is detected by one or other of the sensors, and a corresponding control signal is generated which activates a restoring torque.

Usually, the attitude-control process takes place aboard the satellite, but it is also possible for control signals to be transmitted from earth, based on attitude data obtained from the satellite. Also, where a shift in attitude is desired, an *attitude maneuver* is executed. The control signals needed to achieve this maneuver may be transmitted from an earth station.

Controlling torques may be generated in a number of ways. *Passive attitude control* refers to the use of mechanisms which stabilize the satellite without putting a drain on the satellite's energy supplies; at most, infrequent use is made of these supplies, for example, when thruster jets are impulsed to provide corrective torque. Examples of passive attitude control are *spin stabilization* and *gravity gradient stabilization*. The latter depends on the interaction of the satellite with the gravitational field of the central body and has been used, for example, with the Radio Astronomy Explorer-2 satellite, which was placed in orbit around the moon (Wertz, 1984). For communications satellites, spin stabilization is often used, and this is described in detail in Sec. 7.3.1.

The other form of attitude control is *active control*. With active attitude control, there is no overall stabilizing torque present to resist the disturbance torques. Instead, corrective torques are applied as required in response to disturbance torques. Methods used to generate active control torques include momentum wheels, electromagnetic coils, and mass expulsion devices, such as gas jets and ion thrusters. The electromagnetic coil works on the principle that the earth's magnetic field exerts a torque on a current-carrying coil and that this torque can be controlled through control of the current. However, the method is of use only for satellites relatively close to the earth. The use of momentum wheels is described in more detail in Sec. 7.3.2.

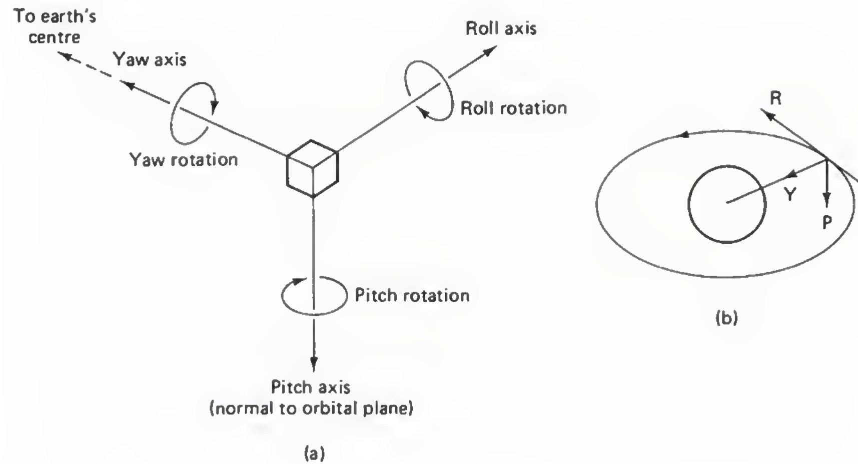


Figure 7.4 (*a*) Roll, pitch, and yaw axes. The yaw axis is directed toward the earth's center, the pitch axis is normal to the orbital plane, and the roll axis is perpendicular to the other two. (*b*) RPY axes for the geostationary orbit. Here, the roll axis is tangential to the orbit and lies along the satellite velocity vector.

The three axes which define a satellite's attitude are its *roll, pitch,* and *yaw* (RPY) axes. These are shown relative to the earth in Fig. 7.4. All three axes pass through the center of gravity of the satellite. For an equatorial orbit, movement of the satellite about the roll axis moves the antenna footprint north and south; movement about the pitch axis moves the footprint east and west; and movement about the yaw axis rotates the antenna footprint.

7.3.1 Spinning satellite stabilization

Spin stabilization may be achieved with cylindrical satellites. The satellite is constructed so that it is mechanically balanced about one particular axis and is then set spinning around this axis. For geostationary satellites, the spin axis is adjusted to be parallel to the N-S axis of the earth, as illustrated in Fig. 7.5. Spin rate is typically in the range of 50 to 100 rev/min. Spin is initiated during the launch phase by means of small gas jets.

In the absence of disturbance torques, the spinning satellite would maintain its correct attitude relative to the earth. Disturbance torques are generated in a number of ways, both external and internal to the satellite. Solar radiation, gravitational gradients, and meteorite impacts are all examples of external forces which can give rise to disturbance torques. Motor-bearing friction and the movement of satellite elements such as the antennas also can give rise to disturbance torques. The

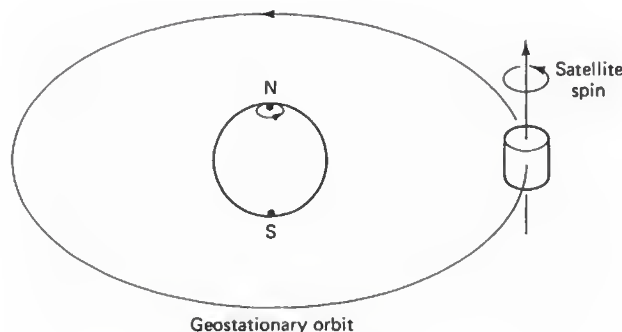


Figure 7.5 Spin stabilization in the geostationary orbit. The spin axis lies along the pitch axis, parallel to the earth's N-S axis.

overall effect is that the spin rate will decrease, and the direction of the angular spin axis will change. Impulse-type thrusters, or jets, can be used to increase the spin rate again and to shift the axis back to its correct N-S orientation. *Nutation*, which is a form of wobbling, can occur as a result of the disturbance torques and/or from misalignment or unbalance of the control jets. This nutation must be damped out by means of energy absorbers known as *nutation dampers*.

Where an omnidirectional antenna is used (e.g., as shown for the INTELSAT I and II satellites in Fig. 1.1), the antenna, which points along the pitch axis, also rotates with the satellite. Where a directional antenna is used, which is more common for communications satellites, the antenna must be despun, giving rise to a dual-spin construction. An electric motor drive is used for despinning the antenna subsystem.

Figure 7.6 shows the Hughes HS 376 satellite in more detail. The antenna subsystem consists of a parabolic reflector and feed horns mounted on the despun shelf, which also carries the communications repeaters (transponders). The antenna feeds can therefore be connected directly to the transponders without the need for radiofrequency (rf) rotary joints, while the complete platform is despun. Of course, control signals and power must be transferred to the despun section, and a mechanical bearing must be provided. The complete assembly for this is known as the *bearing and power transfer assembly* (BAPTA). Figure 7.7 shows a photograph of the internal structure of the HS 376.

Certain dual-spin spacecraft obtain spin stabilization from a spinning flywheel rather than by spinning the satellite itself. These flywheels are termed *momentum wheels*, and their average momentum is referred to as *momentum bias*. Reaction wheels, described in the Sec. 7.3.2, operate at zero momentum bias. In the Intelsat series of satellites, the INTELSAT-VI series spacecraft are spin-stabilized, all the others being 3-axis stabilized (body stabilized) through the use of momentum wheels.

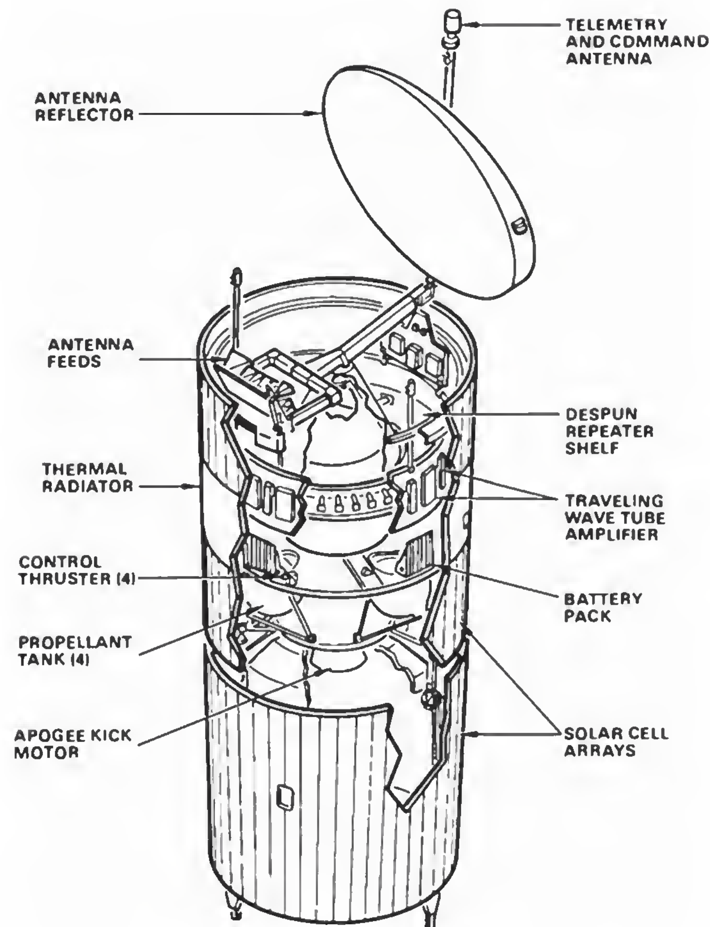


Figure 7.6 HS 376 spacecraft. (*Courtesy of Hughes Aircraft Company Space and Communications Group.*)

7.3.2 Momentum wheel stabilization

In the previous section the gyroscopic effect of a spinning satellite was shown to provide stability for the satellite attitude. Stability also can be achieved by utilizing the gyroscopic effect of a spinning flywheel, and this approach is used in satellites with cube-like bodies (such as shown in Fig. 7.2, and the INTELSAT V type satellites shown in Fig. 1.1). These are known as *body-stabilized* satellites. The complete unit, termed a momentum wheel, consists of a flywheel, the bearing assembly, the casing, and an electric drive motor with associated electronic control circuitry. The flywheel is attached to the rotor, which consists of a

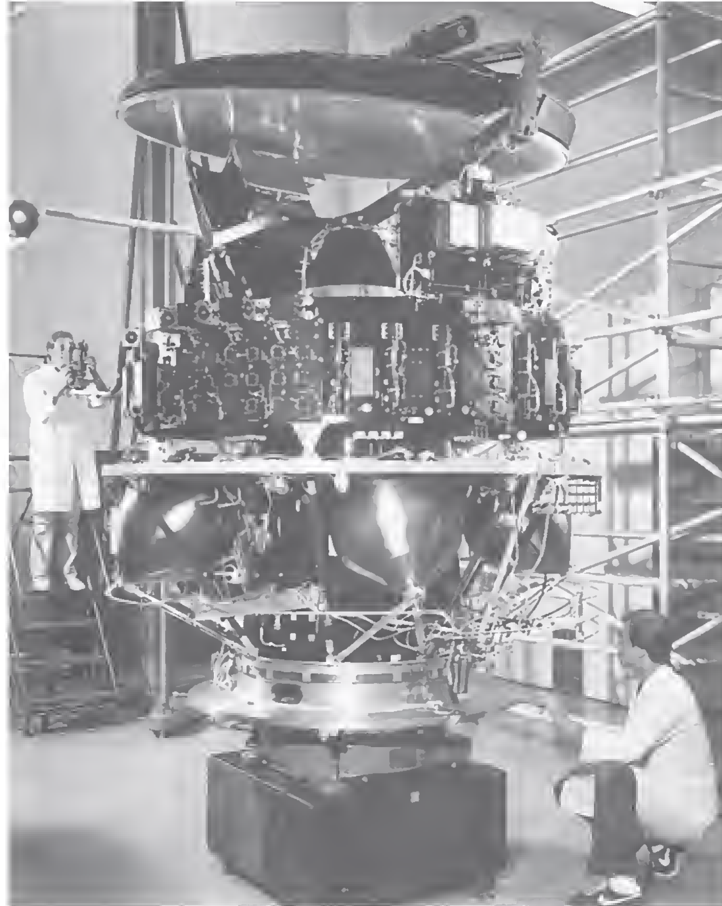


Figure 7.7 Technicians check the alignment of the Telestar 3 communications satellite, shown without its cylindrical panels. The satellite, built for the American Telephone and Telegraph Co., carries both traveling-wave tube and solid-state power amplifiers, as shown on the communications shelf surrounding the center of the spacecraft. The traveling-wave tubes are the cylindrical instruments. (*Courtesy of Hughes Aircraft Company Space and Communications Group.*)

permanent magnet providing the magnetic field for motor action. The stator of the motor is attached to the body of the satellite. Thus the motor provides the coupling between the flywheel and the satellite structure. Speed and torque control of the motor is exercised through the currents fed to the stator. The housing for the momentum wheel is evacuated to protect the wheel from adverse environmental effects, and the bearings have controlled lubrication that lasts over the lifetime of the satellite.

TELDIX manufactures momentum wheels ranging in size from 20, 26, 35, 50, to 60 cm in diameter that are used in a wide variety of satellites. Details of these will be found in Chetty (1991).

The term momentum wheel is usually reserved for wheels that operate at nonzero momentum. This is termed a momentum bias. Such a wheel provides passive stabilization for the yaw and roll axes when the axis of rotation of the wheel lies along the pitch axis, as shown in Fig. 7.8*a*. Control about the pitch axis is achieved by changing the speed of the wheel.

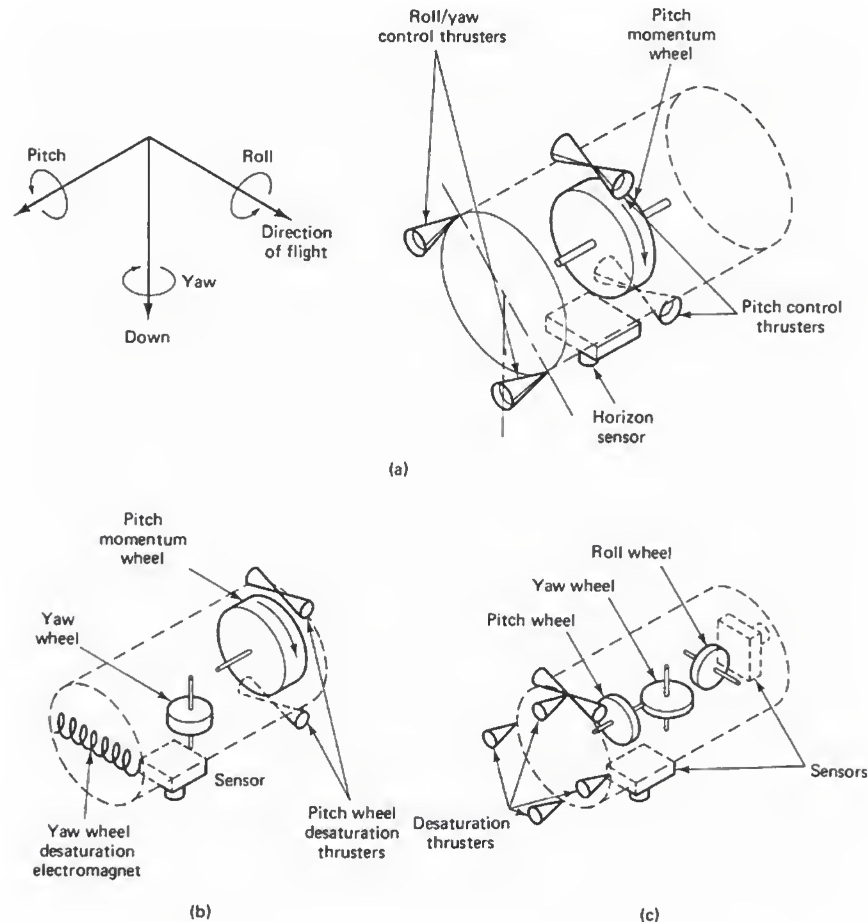


Figure 7.8 Alternative momentum wheel stabilization systems: (*a*) one-wheel, (*b*) two-wheel, (*c*) three-wheel. (*Reprinted with permission from Spacecraft Attitude Determination and Control, edited by James R. Wertz. Copyright, 1984 by D. Reidel Publishing Company, Dordrecht, Holland.*)

When a momentum wheel is operated with zero momentum bias, it is generally referred to as a *reaction wheel*. Reaction wheels are used in three-axis stabilized systems. Here, as the name suggests, each axis is stabilized by a reaction wheel, as shown in Fig. 7.8*c*. Reaction wheels can also be combined with a momentum wheel to provide the control needed (Chetty, 1991). Random and cyclic disturbance torques tend to produce zero momentum on average. However, there will always be some disturbance torques that cause a cumulative increase in wheel momentum, and eventually at some point the wheel *saturates*. In effect, it reaches its maximum allowable angular velocity and can no longer take in any more momentum. Mass expulsion devices are then used to unload the wheel, that is, remove momentum from it (in the same way a brake removes energy from a moving vehicle). Of course, operation of the mass expulsion devices consumes part of the satellite's fuel supply.

7.4 Station Keeping

In addition to having its attitude controlled, it is important that a geostationary satellite be kept in its correct orbital slot. As described in Sec. 2.8.1, the equatorial ellipticity of the earth causes geostationary satellites to drift slowly along the orbit, to one of two stable points, at 75°E and 105°W. To counter this drift, an oppositely directed velocity component is imparted to the satellite by means of jets, which are pulsed once every 2 or 3 weeks. This results in the satellite drifting back through its nominal station position, coming to a stop, and recommencing the drift along the orbit until the jets are pulsed once again. These maneuvers are termed *east-west station-keeping maneuvers*. Satellites in the 6/4-GHz band must be kept within ±0.1° of the designated longitude, and in the 14/12-GHz band, within ±0.05°.

A satellite which is nominally geostationary also will drift in latitude, the main perturbing forces being the gravitational pull of the sun and the moon. These forces cause the inclination to change at a rate of about 0.85°/year. If left uncorrected, the drift would result in a cyclic change in the inclination, going from 0° to 14.67° in 26.6 years (Spilker, 1977) and back to zero, at which the cycle is repeated. To prevent the shift in inclination from exceeding specified limits, jets may be pulsed at the appropriate time to return the inclination to zero. Counteracting jets must be pulsed when the inclination is at zero to halt the change in inclination. These maneuvers are termed *north-south station-keeping maneuvers*, and they are much more expensive in fuel than are east-west station-keeping maneuvers. The north-south station-keeping tolerances are the same as those for east-west station keeping, ±0.1° in the C band and ±0.05° in the Ku band.

Orbital correction is carried out by command from the TT&C earth station, which monitors the satellite position. East-west and north-south station-keeping maneuvers are usually carried out using the same thrusters as are used for attitude control. Figure 7.9 shows typical latitude and longitude variations for the Canadian Anik-C3 satellite which remain after station-keeping corrections are applied.

Satellite altitude also will show variations of about ±0.1 percent of the nominal geostationary height. If, for sake of argument, this is taken as 36,000 km, the total variation in the height is 72 km. A C-band satellite therefore can be anywhere within a box bound by this height and the ±0.1° tolerances on latitude and longitude. Approximating the geostationary radius as 42,164 km (see Sec. 3.1), an angle of 0.2° subtends an arc of approximately 147 km. Thus both the latitude and longitude sides of the box are 147 km. The situation is sketched in Fig. 7.10, which also shows the relative beamwidths of a 30-m and a 5-m antenna. As shown by Eq. (6.33), the −3-dB beamwidth of a 30-m antenna is about 0.12°, and of a 5-m antenna, about 0.7° at 6 GHz. Assuming 38,000 km (typical) for the slant range, the diameter of the 30-m beam at the satellite will be about 80 km. This beam does not encompass the whole

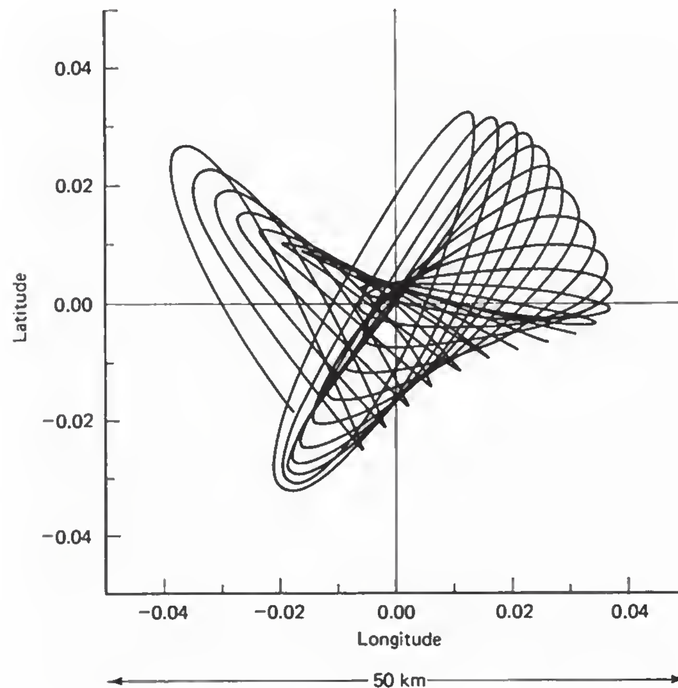


Figure 7.9 Typical satellite motion. (*Courtesy of Telesat, Canada, 1983.*)

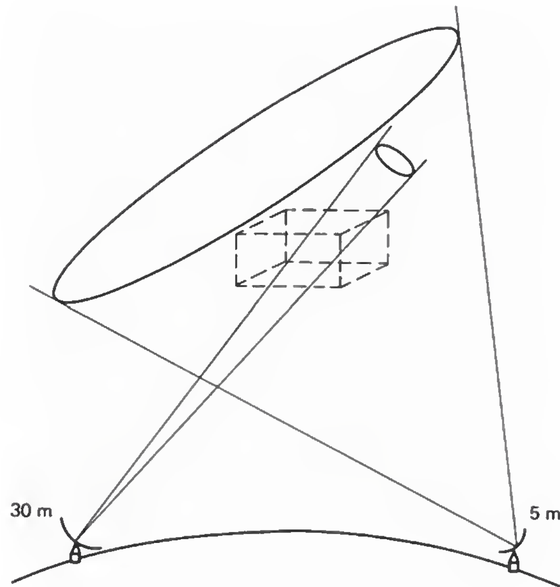


Figure 7.10 The rectangular box shows the positional limits for a satellite in geostationary orbit in relation to beams from a 30-m and a 5-m antenna.

of the box and therefore could miss the satellite. Such narrow-beam antennas therefore must track the satellite.

The diameter of the 5-m antenna beam at the satellite will be about 464 km, and this does encompass the box, so tracking is not required. The positional uncertainty of the satellite also introduces an uncertainty in propagation time, which can be a significant factor in certain types of communications networks.

By placing the satellite in an inclined orbit, the north-south stationkeeping maneuvers may be dispensed with. The savings in weight achieved by not having to carry fuel for these maneuvers allows the communications payload to be increased. The satellite is placed in an inclined orbit of about 2.5° to 3°, in the opposite sense to that produced by drift. Over a period of about half the predicted lifetime of the mission, the orbit will change to equatorial and then continue to increase in inclination. However, this arrangement requires the use of tracking antennas at certain ground stations.

7.5 Thermal Control

Satellites are subject to large thermal gradients, receiving the sun's radiation on one side while the other side faces into space. In addition, thermal radiation from the earth and the earth's *albedo*, which is the

fraction of the radiation falling on earth which is reflected, can be significant for low-altitude earth-orbiting satellites, although it is negligible for geostationary satellites. Equipment in the satellite also generates heat which has to be removed. The most important consideration is that the satellite's equipment should operate as nearly as possible in a stable temperature environment. Various steps are taken to achieve this. Thermal blankets and shields may be used to provide insulation. Radiation mirrors are often used to remove heat from the communications payload. The mirrored thermal radiator for the Hughes HS 376 satellite can be seen in Fig. 7.1 and in Fig. 7.6. These mirrored drums surround the communications equipment shelves in each case and provide good radiation paths for the generated heat to escape into the surrounding space. One advantage of spinning satellites compared with body-stabilized is that the spinning body provides an averaging of the temperature extremes experienced from solar flux and the cold background of deep space.

In order to maintain constant temperature conditions, heaters may be switched on (usually on command from ground) to make up for the heat reduction which occurs when transponders are switched off. The INTELSAT VI satellite used heaters to maintain propulsion thrusters and line temperatures (Pilcher, 1982).

7.6 TT&C Subsystem

The TT&C subsystem performs several routine functions aboard the spacecraft. The telemetry, or telemetering, function could be interpreted as *measurement at a distance.* Specifically, it refers to the overall operation of generating an electrical signal proportional to the quantity being measured and encoding and transmitting this to a distant station, which for the satellite is one of the earth stations. Data which are transmitted as telemetry signals include attitude information such as that obtained from sun and earth sensors; environmental information such as the magnetic field intensity and direction, the frequency of meteorite impact, and so on; and spacecraft information such as temperatures, power supply voltages, and stored-fuel pressure. Certain frequencies have been designated by international agreement for satellite telemetry transmissions. During the transfer and drift orbital phases of the satellite launch, a special channel is used along with an omnidirectional antenna. Once the satellite is on station, one of the normal communications transponders may be used along with its directional antenna, unless some emergency arises which makes it necessary to switch back to the special channel used during the transfer orbit.

Telemetry and command may be thought of as complementary functions. The telemetry subsystem transmits information about the satellite

to the earth station, while the command subsystem receives command signals from the earth station, often in response to telemetered information. The command subsystem demodulates and, if necessary, decodes the command signals and routes these to the appropriate equipment needed to execute the necessary action. Thus attitude changes may be made, communication transponders switched in and out of circuits, antennas redirected, and station-keeping maneuvers carried out on command. It is clearly important to prevent unauthorized commands from being received and decoded, and for this reason, the command signals are often encrypted. *Encrypt* is derived from a Greek word *kryptein*, meaning *to hide*, and represents the process of concealing the command signals in a secure code. This differs from the normal process of encoding which converts characters in the command signal into a code suitable for transmission.

Tracking of the satellite is accomplished by having the satellite transmit beacon signals which are received at the TT&C earth stations. Tracking is obviously important during the transfer and drift orbital phases of the satellite launch. Once it is on station, the position of a geostationary satellite will tend to be shifted as a result of the various disturbing forces, as described previously. Therefore, it is necessary to be able to track the satellite's movement and send correction signals as required. Tracking beacons may be transmitted in the telemetry channel, or by pilot carriers at frequencies in one of the main communications channels, or by special tracking antennas. Satellite range from the ground station is also required from time to time. This can be determined by measurement of the propagation delay of signals especially transmitted for ranging purposes.

It is clear that the telemetry, tracking, and command functions are complex operations which require special ground facilities in addition to the TT&C subsystems aboard the satellite. Figure 7.11 shows in block diagram form the TT&C facilities used by Canadian Telesat for its satellites.

7.7 Transponders

A transponder is the series of interconnected units which forms a single communications channel between the receive and transmit antennas in a communications satellite. Some of the units utilized by a transponder in a given channel may be common to a number of transponders. Thus, although reference may be made to a specific transponder, this must be thought of as an equipment *channel* rather than a single item of equipment.

Before describing in detail the various units of a transponder, the overall frequency arrangement of a typical C-band communications satellite will be examined briefly. The bandwidth allocated for C-band service is 500 MHz, and this is divided into subbands, one for each

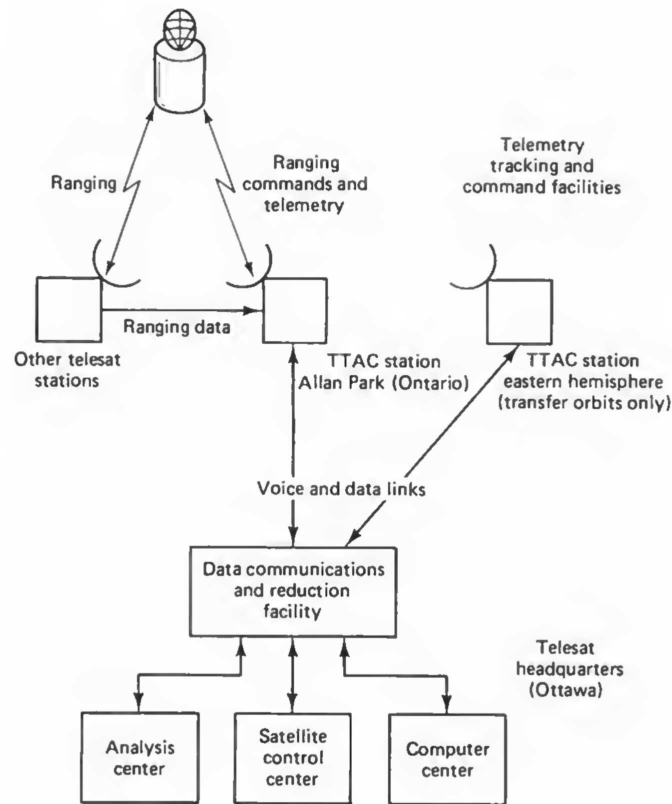


Figure 7.11 Satellite control system. (*Courtesy of Telesat Canada, 1983.*)

transponder. A typical transponder bandwidth is 36 MHz, and allowing for a 4-MHz guardband between transponders, 12 such transponders can be accommodated in the 500-MHz bandwidth. By making use of *polarization isolation*, this number can be doubled. Polarization isolation refers to the fact that carriers, which may be on the same frequency but with opposite senses of polarization, can be isolated from one another by receiving antennas matched to the incoming polarization. With linear polarization, vertically and horizontally polarized carriers can be separated in this way, and with circular polarization, left-hand circular and right-hand circular polarizations can be separated. Because the carriers with opposite senses of polarization may overlap in frequency, this technique is referred to as *frequency reuse*. Figure 7.12 shows part of the frequency and polarization plan for a C-band communications satellite.

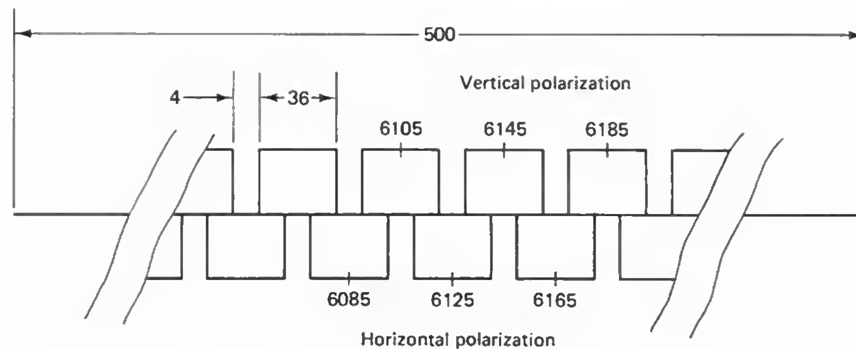


Figure 7.12 Section of an uplink frequency and polarization plan. Numbers refer to frequency in megahertz.

Frequency reuse also may be achieved with spot-beam antennas, and these may be combined with polarization reuse to provide an effective bandwidth of 2000 MHz from the actual bandwidth of 500 MHz.

For one of the polarization groups, Fig. 7.13 shows the channeling scheme for the 12 transponders in more detail. The incoming, or uplink, frequency range is 5.925 to 6.425 GHz. The carriers may be received on one or more antennas, all having the same polarization. The input filter passes the full 500-MHz band to the common receiver while rejecting out-of-band noise and interference such as might be caused by image signals. There will be many modulated carriers within this 500-MHz passband, and all of these are amplified and frequency-converted in the common receiver. The frequency conversion shifts the carriers to the downlink frequency band, which is also 500 MHz wide, extending from 3.7 to 4.2 GHz. At this point the signals are channelized into frequency bands which represent the individual transponder bandwidths.

A transponder may handle one modulated carrier, such as a TV signal, or it may handle a number of separate carriers simultaneously, each modulated by its own telephony or other baseband channel.

7.7.1 The wideband receiver

The wideband receiver is shown in more detail in Fig. 7.14. A duplicate receiver is provided so that if one fails, the other is automatically switched in. The combination is referred to as a *redundant receiver*, meaning that although two are provided, only one is in use at a given time.

The first stage in the receiver is a *low-noise amplifier* (LNA). This amplifier adds little noise to the carrier being amplified, and at the same time it provides sufficient amplification for the carrier to override the higher noise level present in the following mixer stage. In calculations

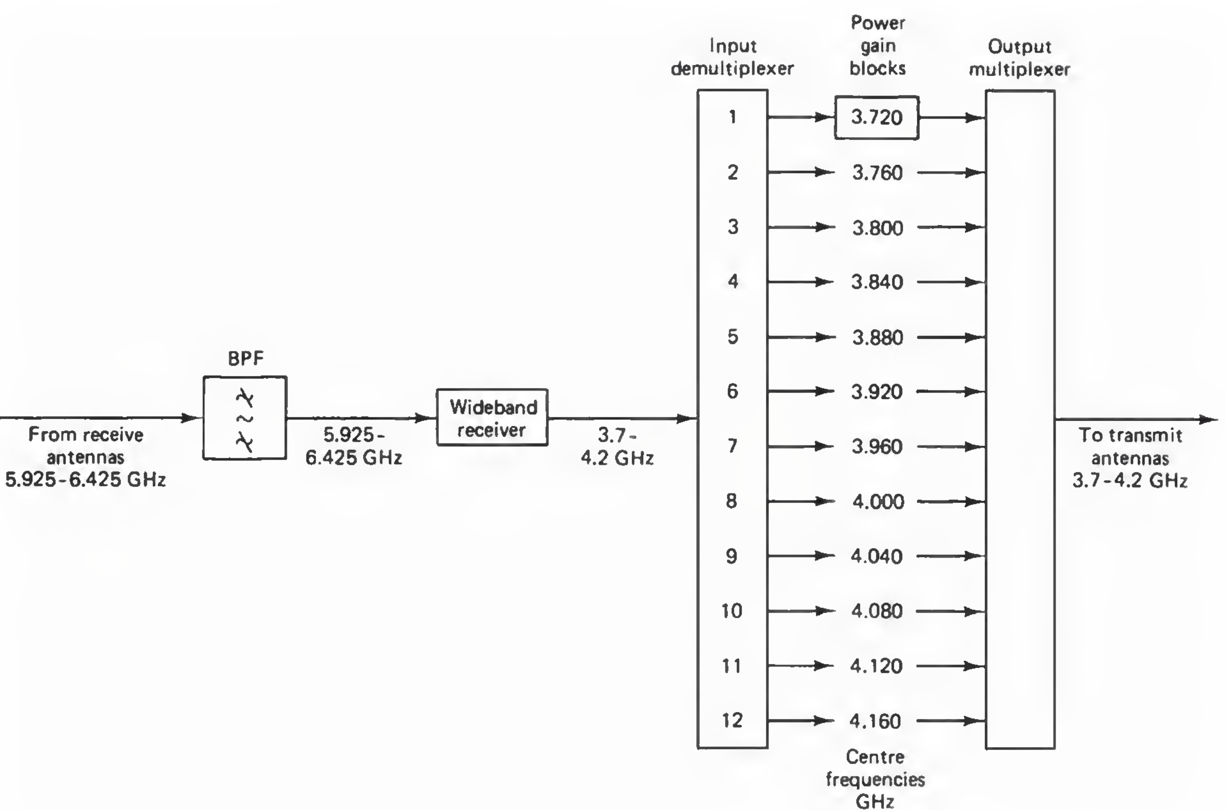


Figure 7.13 Satellite transponder channels. (*Courtesy of CCIR, CCIR Fixed Satellite Services Handbook, final draft 1984.*)

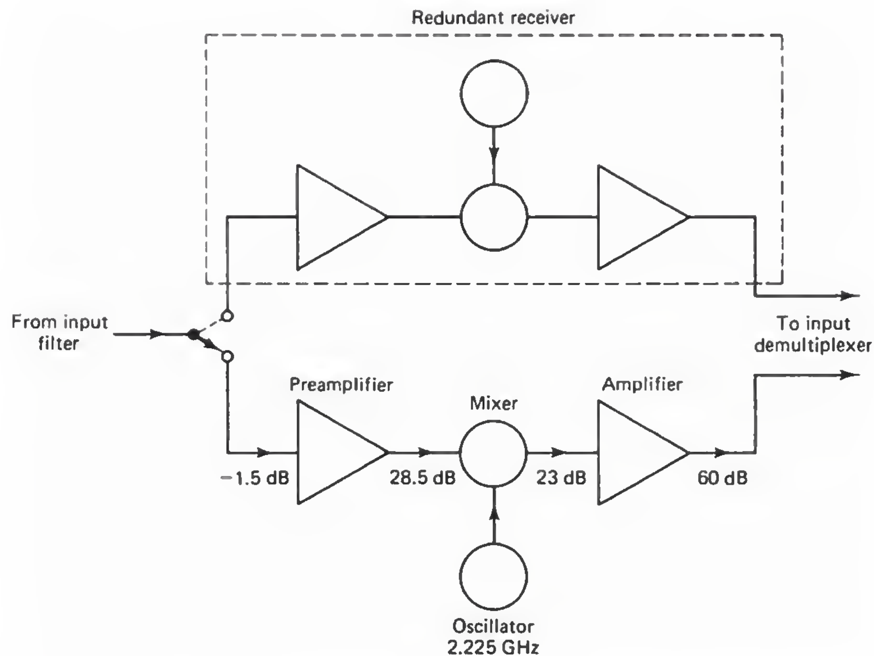


Figure 7.14 Satellite wideband receiver. (*Courtesy of CCIR, CCIR Fixed Satellite Services Handbook, final draft 1984.*)

involving noise, it is usually more convenient to refer all noise levels to the LNA input, where the total receiver noise may be expressed in terms of an equivalent noise temperature. In a well-designed receiver, the equivalent noise temperature referred to the LNA input is basically that of the LNA alone. The overall noise temperature must take into account the noise added from the antenna, and these calculations are presented in detail in Chap. 12. The equivalent noise temperature of a satellite receiver may be on the order of a few hundred kelvins.

The LNA feeds into a mixer stage, which also requires a *local oscillator* (LO) signal for the frequency-conversion process. The power drive from the LO to the mixer input is about 10 dBm. The oscillator frequency must be highly stable and have low-phase noise. A second amplifier follows the mixer stage to provide an overall receiver gain of about 60 dB. The signal levels in decibels referred to the input are shown in Fig. 7.14 (CCIR, 1984). Splitting the gain between the preamplifier at 6 GHz and the second amplifier at 4 GHz prevents oscillation, which might occur if all the gain were to be provided at the same frequency.

The wideband receiver utilizes only solid-state active devices. In some designs, tunnel-diode amplifiers have been used for the preamplifier at 6 GHz in 6/4-GHz transponders and for the parametric amplifiers at 14 GHz

in 14/12-GHz transponders. With advances in *field-effect transistor* (FET) technology, FET amplifiers, which offer equal or better performance, are now available for both bands. Diode mixer stages are used. The amplifier following the mixer may utilize *bipolar junction transistors* (BJTs) at 4 GHz and FETs at 12 GHz, or FETs may in fact be used in both bands.

7.7.2 The input demultiplexer

The input demultiplexer separates the broadband input, covering the frequency range 3.7 to 4.2 GHz, into the transponder frequency channels. Referring to Fig. 7.13, for example, the separate channels labeled 1 through 12 are shown in detail in Fig. 7.15. The channels are usually arranged in even-numbered and odd-numbered groups. This provides greater frequency separation between adjacent channels in a group, which reduces adjacent channel interference. The output from the receiver is fed to a power splitter, which in turn feeds the two separate chains of circulators. The full broadband signal is transmitted along each chain, and the channelizing is achieved by means of channel filters connected to each circulator, as shown in Fig. 7.15. The channel numbers correspond to those shown in Fig. 7.13. Each filter has a bandwidth of 36 MHz and is tuned to the appropriate center frequency, as shown in Fig. 7.13. Although there are considerable losses in the demultiplexer, these are easily made up in the overall gain for the transponder channels.

7.7.3 The power amplifier

A separate power amplifier provides the output power for each transponder channel. As shown in Fig. 7.16, each power amplifier is preceded by an input attenuator. This is necessary to permit the input drive to each power amplifier to be adjusted to the desired level. The attenuator has a fixed section and a variable section. The fixed attenuation is needed to balance out variations in the input attenuation so that each transponder channel has the same nominal attenuation, the necessary adjustments being made during assembly. The variable attenuation is needed to set the level as required for different types of service (an example being the requirement for input power backoff discussed later). Because this variable attenuator adjustment is an operational requirement, it must be under the control of the ground TT&C station.

Traveling-wave tube amplifiers (TWTAs) are widely used in transponders to provide the final output power required to the transmit antenna. Figure 7.17 shows the schematic of a *traveling wave tube* (TWT) and its power supplies. In the TWT, an electron-beam gun assembly consisting of a heater, a cathode, and focusing electrodes is used to form an electron beam. A magnetic field is required to confine the beam to travel along the inside of a wire helix. For high-power tubes, such as might be

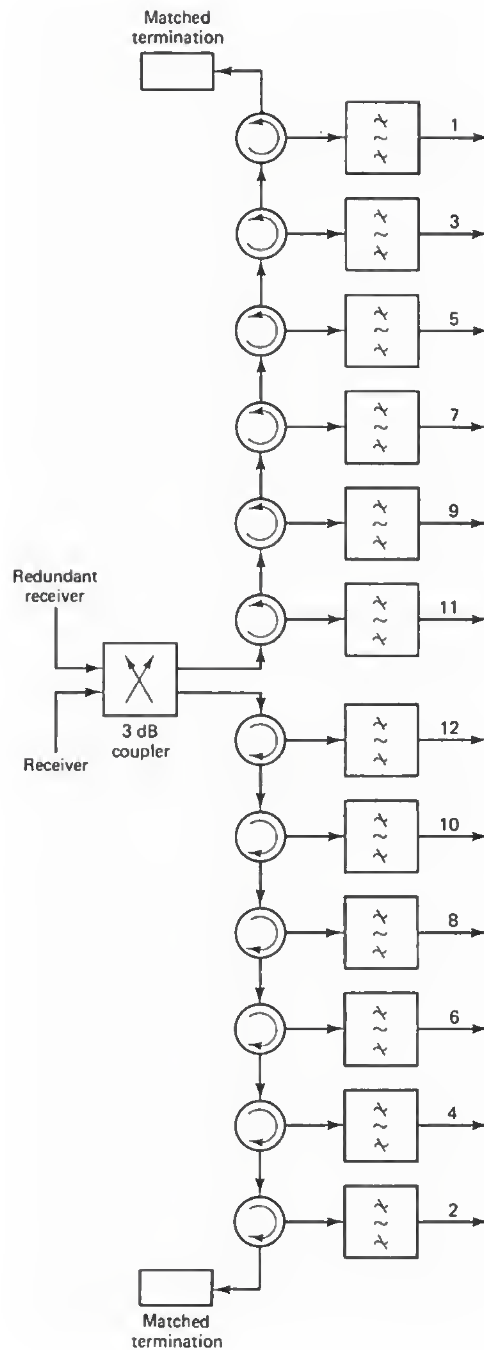


Figure 7.15 Input demultiplexer. (*Courtesy of CCIR, CCIR Fixed Satellite Services Handbook, final draft 1984.*)

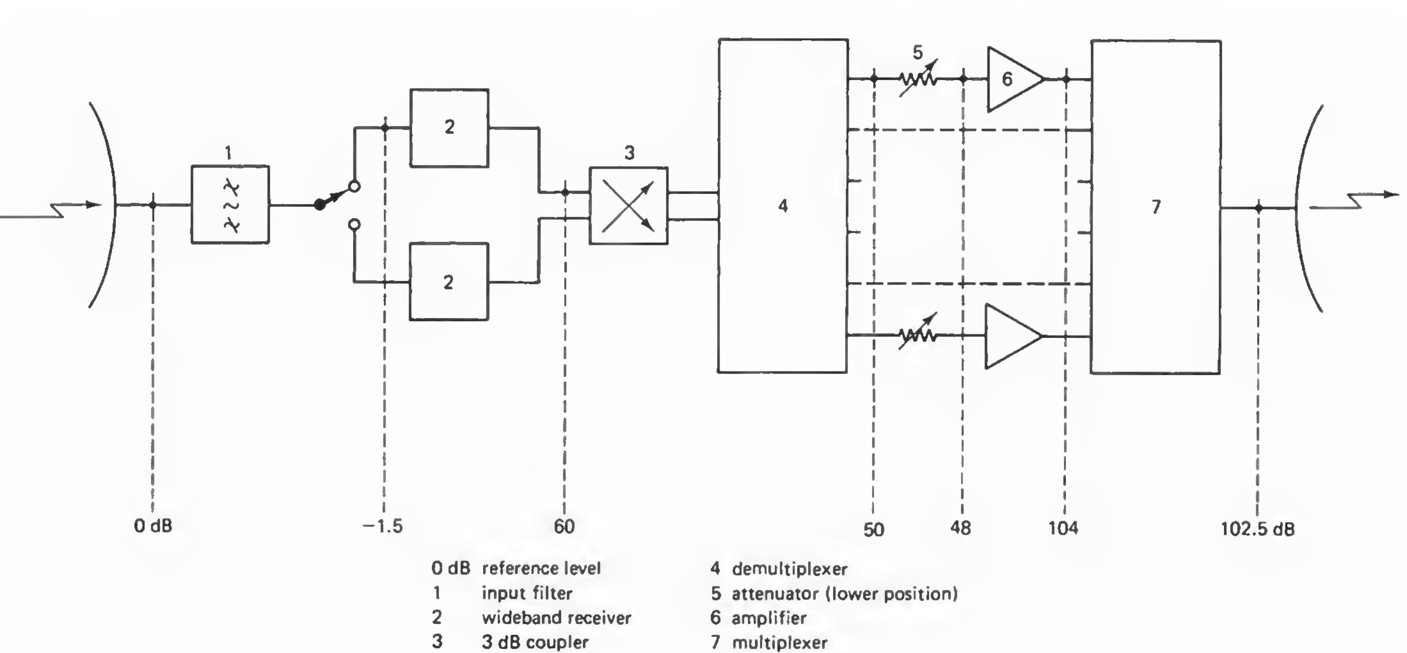


Figure 7.16 Typical diagram of the relative levels in a transponder. (*Courtesy of CCIR, CCIR Fixed Satellite Services Handbook, final draft 1984.*)

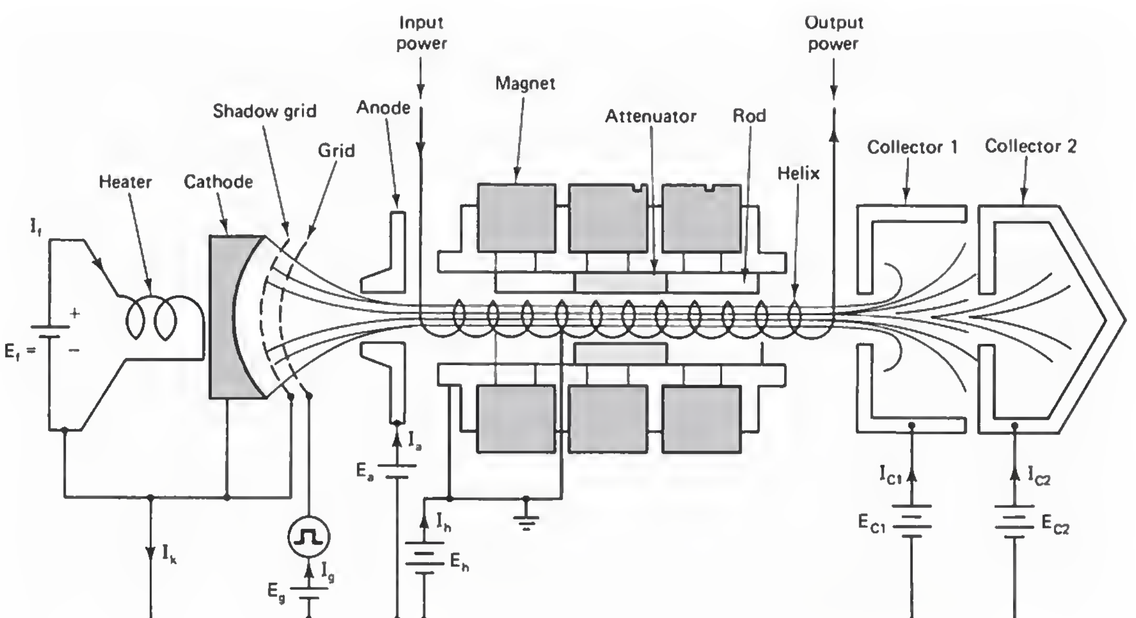


Figure 7.17 Schematic of a TWT and power supplies. (*Courtesy of Hughes TWT and TWTA Handbook; courtesy of Hughes Aircraft Company, Electron Dynamics Division, Torrance, CA.*)

used in ground stations, the magnetic field can be provided by means of a solenoid and dc power supply. The comparatively large size and high power consumption of solenoids make them unsuitable for use aboard satellites, and lower-power TWTs are used which employ permanent-magnet focusing.

The rf signal to be amplified is coupled into the helix at the end nearest the cathode and sets up a traveling wave along the helix. The electric field of the wave will have a component along the axis of the helix. In some regions, this field will decelerate the electrons in the beam, and in others it will accelerate them so that electron bunching occurs along the beam. The average beam velocity, which is determined by the dc potential on the tube collector, is kept slightly greater than the phase velocity of the wave along the helix. Under these conditions, an energy transfer takes place, kinetic energy in the beam being converted to potential energy in the wave. The wave actually will travel around the helical path at close to the speed of light, but it is the axial component of wave velocity which interacts with the electron beam. This component is less than the velocity of light approximately in the ratio of helix pitch to circumference. Because of this effective reduction in phase velocity, the helix is referred to as a *slowwave structure.*

The advantage of the TWT over other types of tube amplifiers is that it can provide amplification over a very wide bandwidth. Input levels to the TWT must be carefully controlled, however, to minimize the effects of certain forms of distortion. The worst of these result from the nonlinear transfer characteristic of the TWT, illustrated in Fig. 7.18.

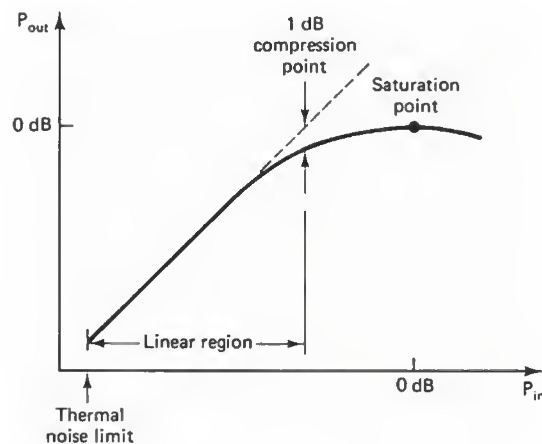


Figure 7.18 Power transfer characteristics of a TWT. The saturation point is used as 0-dB reference for both input and output.

At low-input powers, the output-input power relationship is linear; that is, a given decibel change in input power will produce the same decibel change in output power. At higher power inputs, the output power saturates, the point of maximum power output being known as the *saturation point*. The saturation point is a very convenient reference point, and input and output quantities are usually referred to it. The linear region of the TWT is defined as the region bound by the thermal noise limit at the low end and by what is termed the *1-dB compression point* at the upper end. This is the point where the actual transfer curve drops 1 dB below the extrapolated straight line, as shown in Fig. 7.18. The selection of the operating point on the transfer characteristic will be considered in more detail shortly, but first the phase characteristics will be described. The absolute time delay between input and output signals at a fixed input level is generally not significant. However, at higher input levels, where more of the beam energy is converted to output power, the average beam velocity is reduced, and therefore, the delay time is increased. Since phase delay is directly proportional to time delay, this results in a phase shift which varies with input level. Denoting the phase shift at saturation by θ_S and in general by θ, the phase difference relative to saturation is $\theta - \theta_S$. This is plotted in Fig. 7.19 as a function of input power. Thus, if the input signal power level changes, phase modulation will result, this being termed *AM/PM conversion*. The slope of the phase shift characteristic gives the phase modulation coefficient, in degrees per decibel. The curve of the slope as a function of input power is also sketched in Fig. 7.19.

Frequency modulation (FM) is usually employed in analog satellite communications circuits. However, unwanted *amplitude modulation* (AM) can occur from the filtering which takes place prior to the TWT input. The AM process converts the unwanted amplitude modulation to *phase modulation* (PM), which appears as noise on the FM carrier.

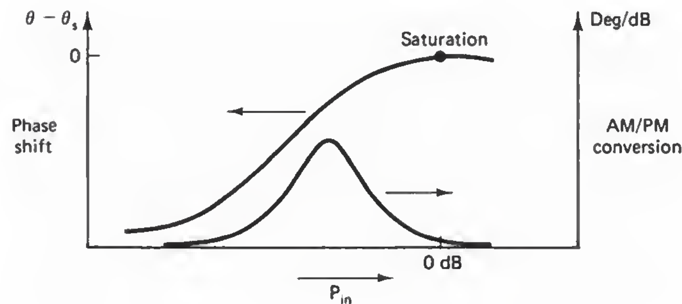


Figure 7.19 Phase characteristics for a TWT. θ is the input-to-output phase shift, and θ_S is the value at saturation. The AM/PM curve is derived from the slope of the phase shift curve.

Where only a single carrier is present, it may be passed through a *hard limiter* before being amplified in the TWT. The hard limiter is a circuit which clips the carrier amplitude close to the zero baseline to remove any amplitude modulation. The FM is preserved in the zero crossover points and is not affected by the limiting.

A TWT also may be called on to amplify two or more carriers simultaneously, this being referred to as *multicarrier operation*. The AM/PM conversion is then a complicated function of carrier amplitudes, but in addition, the nonlinear transfer characteristic introduces a more serious form of distortion known as *intermodulation distortion*. The nonlinear transfer characteristic may be expressed as a Taylor series expansion which relates input and output voltages:

$$e_0 = ae_i + be_i^2 + ce_i^3 + \cdots \tag{7.1}$$

Here, *a, b, c*, and so on are coefficients which depend on the transfer characteristic, e_0 is the output voltage, and e_i is the input voltage, which consists of the sum of the individual carriers. The *third-order term* is ce_i^3. This and higher-order odd-power terms give rise to intermodulation products, but usually only the third-order contribution is significant. Suppose multiple carriers are present, separated from one another by Δf, as shown in Fig. 7.20. Considering specifically the carriers at frequencies f_1 and f_2, these will give rise to frequencies $2f_2 - f_1$ and $2f_1 - f_2$ as a result of the third-order term. (This is demonstrated in App. E.)

Because $f_2 - f_1 = \Delta f$, these two intermodulation products can be written as $f_2 + \Delta f$ and $f_1 - \Delta f$, respectively. Thus the intermodulation products fall on the neighboring carrier frequencies as shown in Fig. 7.20. Similar intermodulation products will arise from other carrier pairs, and when the carriers are modulated the intermodulation distortion appears as noise across the transponder frequency band. This intermodulation noise is discussed further in Sec. 12.10.

In order to reduce the intermodulation distortion, the operating point of the TWT must be shifted closer to the linear portion of the curve, the

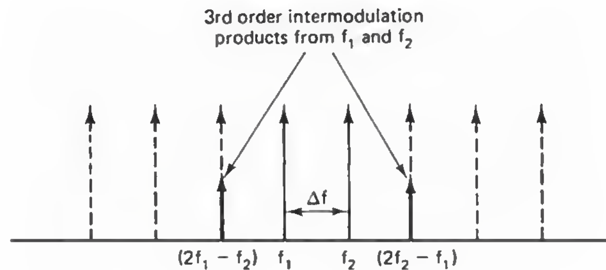


Figure 7.20 Third-order intermodulation products.

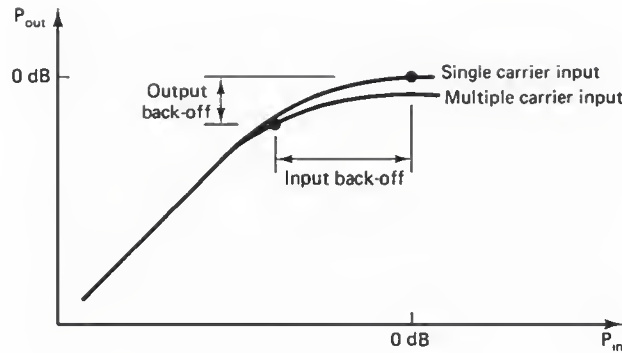


Figure 7.21 Transfer curve for a single carrier and for one carrier of a multiple-carrier input. Backoff for multiple-carrier operation is relative to saturation for single-carrier input.

reduction in input power being referred to as *input backoff*. When multiple carriers are present, the power output around saturation, for any one carrier, is less than that achieved with single-carrier operation. This is illustrated by the transfer curves of Fig. 7.21. The input backoff is the difference in decibels between the carrier input at the operating point and the saturation input which would be required for single-carrier operation. The output backoff is the corresponding drop in output power. Backoff values are always stated in decibels relative to the saturation point. As a rule of thumb, output backoff is about 5 dB less than input backoff. The need to incorporate backoff significantly reduces the channel capacity of a satellite link because of the reduced carrier-to-noise ratio received at the ground station. Allowance for backoff in the link budget calculations is dealt with in Secs. 12.7.2 and 12.8.1.

7.8 The Antenna Subsystem

The antennas carried aboard a satellite provide the dual functions of receiving the uplink and transmitting the downlink signals. They range from dipole-type antennas where omnidirectional characteristics are required to the highly directional antennas required for telecommunications purposes and TV relay and broadcast. Parts of the antenna structures for the HS 376 and HS 601 satellites can be seen in Figs. 7.1, 7.2, and 7.7.

Directional beams are usually produced by means of reflector-type antennas—the paraboloidal reflector being the most common. As shown in Chap. 6, the gain of the paraboloidal reflector, relative to an isotropic radiator, is given by Eq. (6.32)

$$G = \eta_I \left(\frac{\pi D}{\lambda} \right)^2$$

where λ is the wavelength of the signal, D is the reflector diameter, and η_I is the aperture efficiency. A typical value for η_I is 0.55. The −3-dB beamwidth is given approximately by Eq. (6.33) as

$$\theta_{3\text{dB}} \cong 70\frac{\lambda}{D} \text{ degrees}$$

The ratio D/λ is seen to be the key factor in these equations, the gain being directly proportional to $(D/\lambda)^2$ and the beamwidth inversely proportional to D/λ. Hence the gain can be increased and the beamwidth made narrower by increasing the reflector size or decreasing the wavelength. In comparing C-band and Ku-band, the largest reflectors are those for the 6/4-GHz band. Comparable performance can be obtained with considerably smaller reflectors in the 14/12-GHz band. Satellites used for mobile services in the L-band employ much larger antennas (with reflector areas in the order of 100 m^2 to 200 m^2) as described in Chap. 17.

Figure 7.22 shows the antenna subsystem of the INTELSAT VI satellite (Johnston and Thompson, 1982). This provides a good illustration of the level of complexity which has been reached in large communications satellites. The largest reflectors are for the 6/4-GHz hemisphere and zone coverages, as illustrated in Fig. 7.23. These are fed from horn arrays, and various groups of horns can be excited to produce the beam shape required. As can be seen, separate arrays are used for transmit and receive. Each array has 146 dual-polarization horns. In the 14/11-GHz band, circular reflectors are used to provide spot beams, one for east and one for west, also shown in Fig. 7.23. These beams are fully steerable. Each spot is fed by a single horn which is used for both transmit and receive.

Wide beams for global coverage are produced by simple horn antennas at 6/4 GHz. These horns beam the signal directly to the earth without the use of reflectors. Also as shown in Fig. 7.22, a simple biconical dipole antenna is used for the tracking and control signals. The complete antenna platform and the communications payload are despun as described in Sec. 7.3 to keep the antennas pointing to their correct locations on earth.

The same feed horn may be used to transmit and receive carriers with the same polarization. The transmit and receive signals are separated in a device known as a *diplexer*, and the separation is further aided by means of frequency filtering. Polarization discrimination also may be used to separate the transmit and receive signals using the same feed horn. For example, the horn may be used to transmit horizontally polarized waves in the downlink frequency band, while simultaneously receiving vertically polarized waves in the uplink frequency

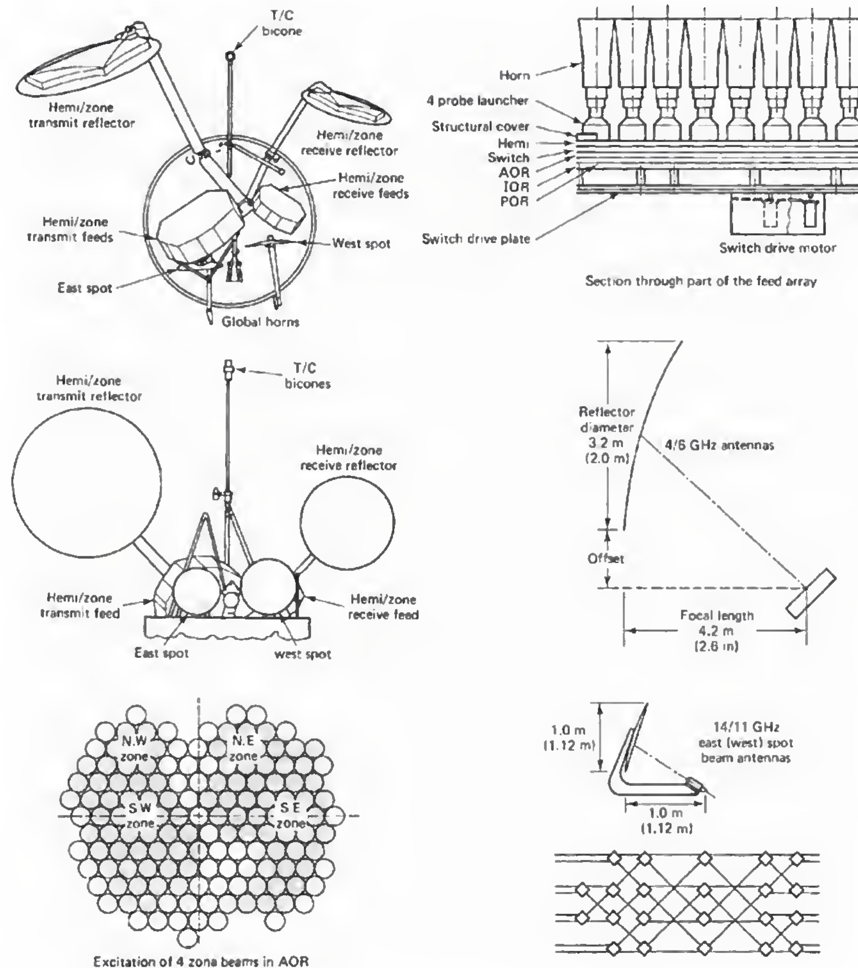


Figure 7.22 The antenna subsystem for the INTELSAT VI satellite. (*Courtesy of Johnston and Thompson, 1982, with permission.*)

band. The polarization separation takes place in a device known as an *orthocoupler*, or *orthogonal mode transducer* (OMT). Separate horns also may be used for the transmit and receive functions, with both horns using the same reflector.

7.9 Morelos and Satmex 5

Figure 7.24 shows the communications subsystem of the Mexican satellite Morelos. Two such satellites were launched, Morelos A in June and Morelos B in November 1985. The satellites are from the Hughes 376

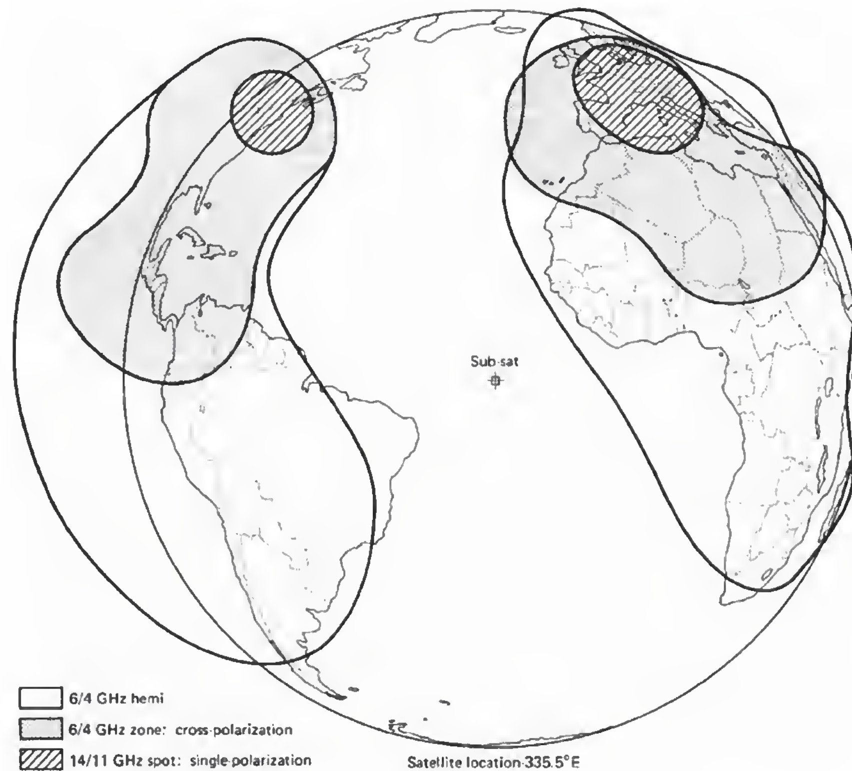


Figure 7.23 INTELSAT V Atlantic satellite transmit capabilities. (Note: The 14/11-GHz spot beams are steerable and may be moved to meet traffic requirements as they develop.) (*Courtesy of Intelsat Document BG-28-72E M/6/77, with permission.*)

spacecraft series. These satellites had a predicted mission life of 9 years, Morelos-A being retired in 1994, and Morelos-B was scheduled to remain in operation until 1998. The payload carried on Morelos illustrates what is referred to as a *hybrid*, or *dual-band, payload* because it carried C-band and K-band transponders. In the C band, Morelos provided 12 narrowband channels, each 36-MHz wide, and 6 wideband channels, each 72-MHz wide. In the K band it provided four channels, each 108-MHz wide. The 36-MHz channels used 7-W TWTAs with 14-for-12 redundancy. This method of stating redundancy simply means that 12 redundant units are available for 14 in-service units. The 72-MHz channels used 10.5-W TWTAs with 8-for-6 redundancy.

The four K-band repeaters used six 20-W TWTAs with 6-for-4 redundancy. The receivers were solid-state designs, with a 4-for-2 redundancy for the C-band receivers and 2-for-1 redundancy for the K-band receivers.

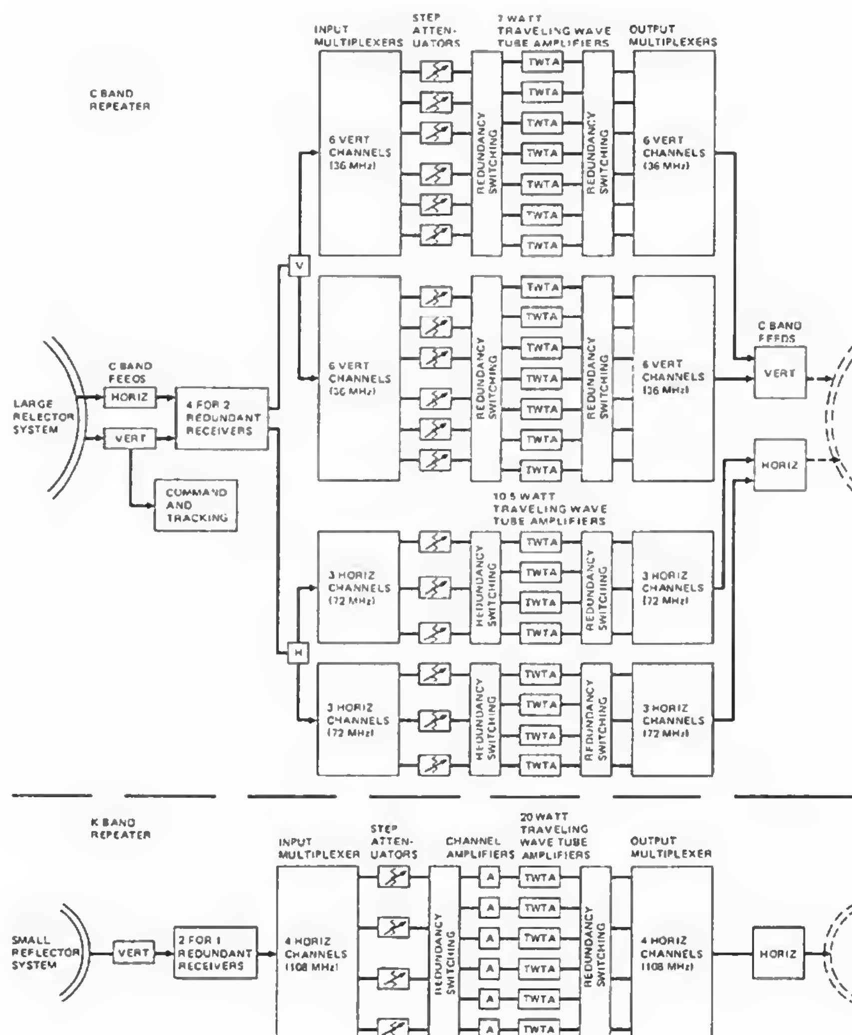


Figure 7.24 Communications subsystem functional diagram for Morelos. (*Courtesy of Hughes Aircraft Company Space and Communications Group.*)

As mentioned, the satellites are part of the Hughes 376 series, illustrated in Figs. 7.1 and 7.6. A 180-cm-diameter circular reflector is used for the C band. This forms part of a dual-polarization antenna, with separate C-band feeds for horizontal and vertical polarizations. Morelos provided the C-band footprint pattern shown in Fig. 7.25*a*.

The Morelos K-band reflector was elliptical in shape, with axes measuring 150 by 91 cm. It had its own feed array, producing a footprint which closely matched the contours of the Mexican land mass, as shown

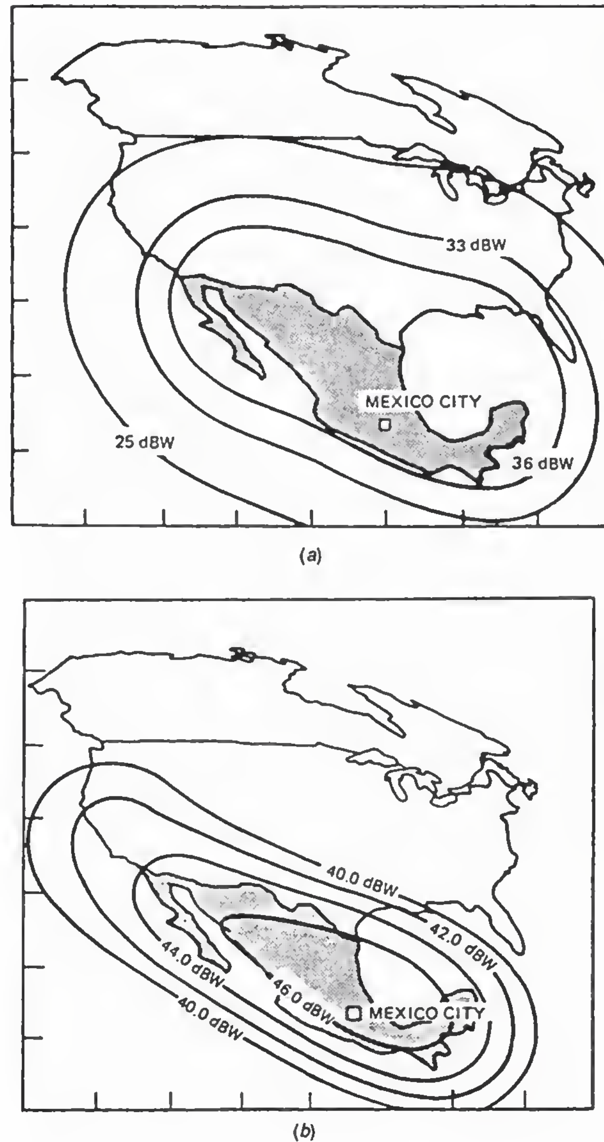


Figure 7.25 (a) C-band and (b) K-band transmit coverage for Morelos. (*Courtesy of Hughes Aircraft Company Space and Communications Group.*)

in Fig. 7.25*b*. The K-band reflector was tied to the C-band reflector, and onboard tracking of a C-band beacon transmitted from the Tulancingo TT&C station ensured precise pointing of the antennas.

On December 5, 1998 the SATMEX-5 was launched, with an expected life of 15 years. SATMEX-5 is a Hughes 601HP, which is a high-powered

version of the Hughes 601 satellite illustrated in Fig. 7.2. The antenna footprints have been expanded to cover the whole American continent. In the Ku-band coverage area it will be possible to receive *direct-to-home* (DTH) television broadcasts on antennas 60 cm diameter or less. SATMEX-5 occupies the 116.8°W slot on the geostationary orbit.

7.10 Anik-Satellites

The Anik series of satellites are designed to provide communications services in Canada as well as cross-border services with the United States. Early satellites such as Anik C and D were both Boeing 376 models. Anik-E was the first of the body-stabilized satellites used in the series, a dual-band satellite which had an equivalent capacity of 56 television channels, or more than 50,000 telephone circuits. Attitude control was of the momentum-bias, three-axes-stabilized type, and solar sails were used to provide power, the capacity being 3450 W at summer solstice and 3700 W at vernal equinox. Four NiH_2 batteries provided power during eclipse. Although the Anik-E has been superseded by the Anik-F series, the Anik-E configuration as shown in Fig. 7.26 provides

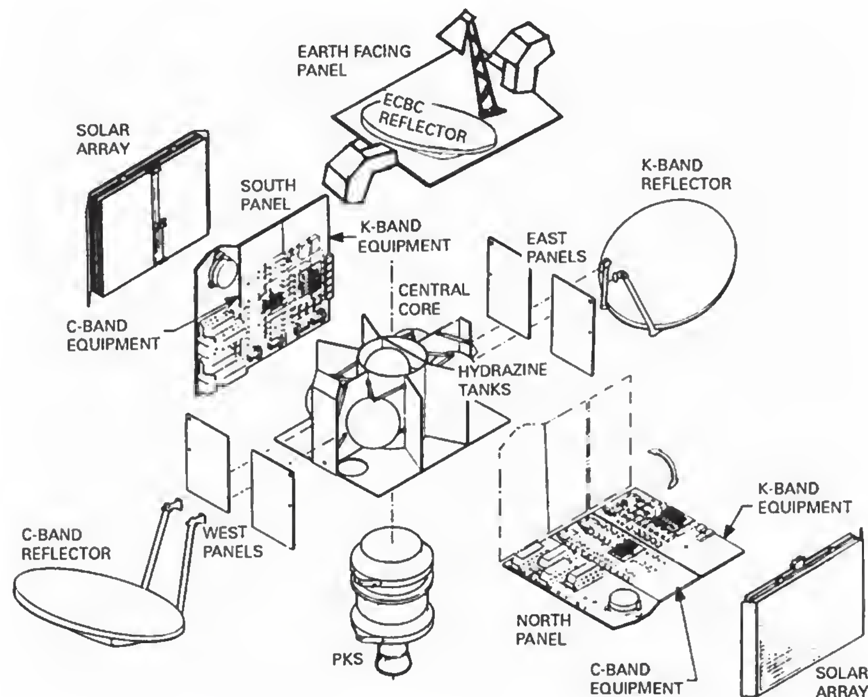


Figure 7.26 Anik-E spacecraft configuration. (*Courtesy of Telesat Canada.*)

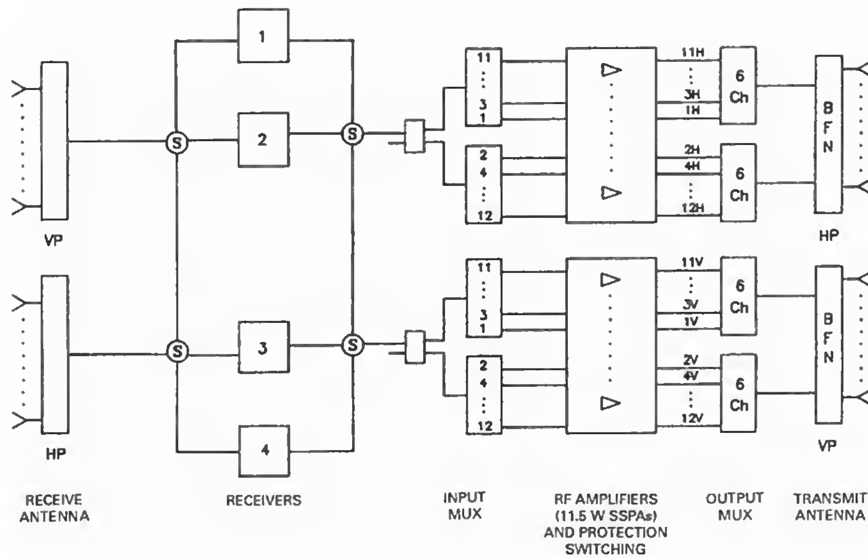


Figure 7.27 Anik-E C-band transponder functional block diagram. (*Courtesy of Telesat Canada.*)

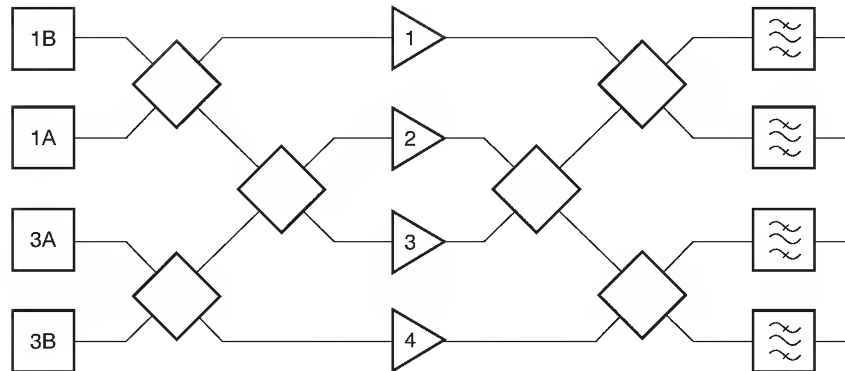
a good illustration of the body-stabilized type of satellite, and Fig. 7.27 shows a typical C-band transponder set-up. This is seen to use *solid-state power amplifiers* (SSPAs) which offer significant improvement in reliability and weight saving over traveling-wave tube amplifiers. The antennas are fed through a *broadband feeder network* (BFN) to illuminate the large reflectors shown in Fig. 7.26. National, as distinct from regional, coverage is provided at C band.

The TWTAs aboard a satellite also may be switched to provide redundancy, as illustrated in Fig. 7.28. The scheme shown is termed a *4-for-2 redundancy*, meaning that four channels are provided with two redundant amplifiers. For example, examination of the table in Fig. 7.28 shows that channel 1A has amplifier 2 as its primary amplifier, and amplifiers 1 and 3 can be switched in as backup amplifiers by ground command.

7.11 Advanced Tiros-N Spacecraft

Tiros is an acronym for *Television and Infra-Red Observational Satellite*. As described in Chap. 1, Tiros is a polar-orbiting satellite, the primary mission of which is to gather and transmit earth environmental data down to its earth stations. Although its payload differs fundamentally from the communications-relay-type payload, much of the bus equipment is similar.

Table 1.7 lists the *National Oceanographic and Atmospheric Administration* (NOAA) spacecraft used in the Advanced TIROS-N (ATN)

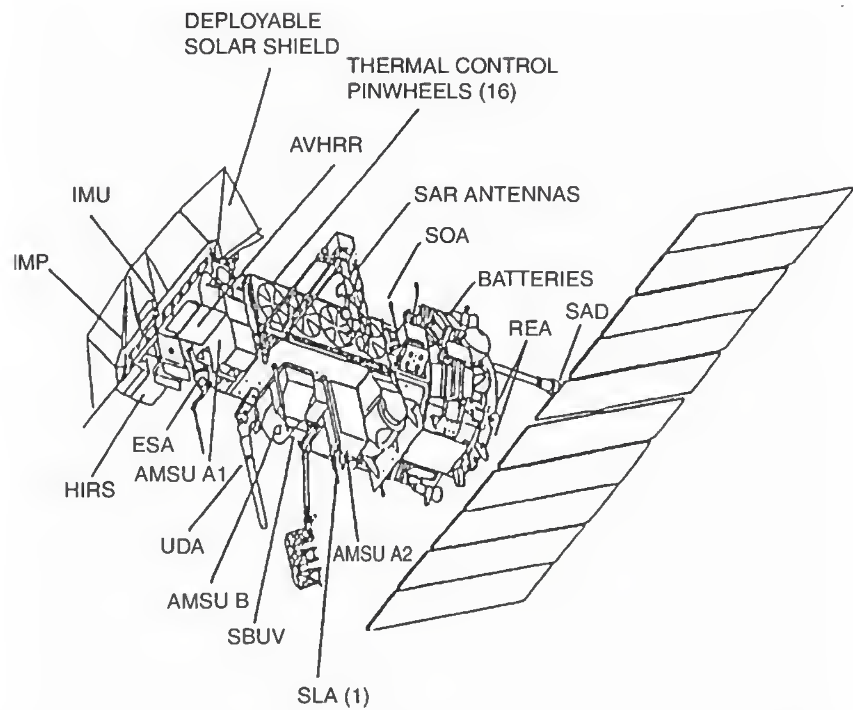


CHANNEL	1A	3A	1B	3B
TWTA				
PRIMARY	2	3	1	4
BACKUP	1 or 3	2 or 4	2 or 3	2 or 3

Figure 7.28 A 4-for-2 redundancy switching arrangement. (*Courtesy of Telesat Canada, 1983.*)

program. The general features of these spacecraft are described in Schwalb (1982*a*, 1982*b*), and current information can be obtained at the NOAA Web site http://www.noaa.gov/. The main features of the NOAA KLM spacecraft are shown in Fig. 7.29, and the physical and orbital characteristics are given in Table 7.1.

Three Ni-Cd batteries supply power while the spacecraft is in darkness. The relatively short lifetime of these spacecraft results largely from the effects of atmospheric drag present at the low orbital altitudes. Attitude control of the NOAA spacecraft is achieved through the use of three reaction wheels similar to the arrangement shown in Fig. 7.8. A fourth, spare wheel is carried, angled at 54.7° to each of the three orthogonal axes. The spare reaction wheel is normally idle but is activated in the event of failure of any of the other wheels. The 54.7° angle permits its torque to be resolved into components along each of the three main axes. As can be seen from Fig. 7.29, the antennas are omnidirectional, but attitude control is needed to maintain directivity for the earth sensors. These must be maintained within ±0.2° of the local geographic reference (Schwalb, 1982*a*).



LEGEND	
AMSU	ADVANCED MICROWAVE SOUNDING UNIT
AVHRR	ADVANCED VERY HIGH RESOLUTION RADIOMETER
ESA	EARTH SENSOR ASSEMBLY
HIRS	HIGH RESOLUTION INFRARED SOUNDER
IMP	INSTRUMENT MOUNTING PLATFORM
IMU	INERTIAL MEASUREMENT UNIT
MHS	MICROWAVE HUMIDITY SOUNDER
REA	REACTION ENGINE ASSEMBLY
SAD	SOLAR ARRAY DRIVE
SAR	SEARCH AND RESCUE
SBUV	SOLAR BACKSCATTER ULTRAVIOLET SOUNDING SPECTRAL RADIOMETER
SOA	S-BAND OMNI ANTENNA
SLA	SEARCH AND RESCUE TRANSMITTING ANTENNA (L BAND)
UDA	ULTRA HIGH FREQUENCY DATA COLLECTION SYSTEM ANTENNA
VRA	VERY HIGH FREQUENCY REALTIME ANTENNA

Figure 7.29 NOAA-KLM spacecraft configuration. (*Courtesy of NOAA National Environmental Satellite, Data, and Information Service.*)

TABLE 7.1 NOAA-15 Characteristics

Main body	4.2 m (13.75 ft) long, 1.88 m (6.2 ft) diameter
Solar array	2.73 m (8.96 ft) by 6.14 m (20.16 ft)
Weight at liftoff	2231.7 kg (4920 lb) including 756.7 kg of expendable fuel
Launch vehicle	Lockheed Martin Titan II
Orbital information	Type: Sun synchronous Altitude: 833 km Period: 101.2 minutes Inclination: 98.70°

SOURCE: Data obtained from http://140.90.207.25:8080/EBB/ml/genlsatl.html.

7.12 Problems and Exercises

7.1. Describe the TT&C facilities of a satellite communications system. Are these facilities part of the space segment or part of the ground segment of the system?

7.2. Explain why some satellites employ cylindrical solar arrays, whereas others employ solar-sail arrays for the production of primary power. State the typical power output to be expected from each type. Why is it necessary for satellites to carry batteries in addition to solar-cell arrays?

7.3. Explain what is meant by satellite *attitude*, and briefly describe two forms of attitude control.

7.4. Define and explain the terms *roll, pitch*, and *yaw*.

7.5. Explain what is meant by the term *despun antenna*, and briefly describe one way in which the despinning is achieved.

7.6. Briefly describe the three-axis method of satellite stabilization.

7.7. Describe the east-west and north-south station-keeping maneuvers required in satellite station keeping. What are the angular tolerances in station keeping that must be achieved?

7.8. Referring to Fig. 7.10 and the accompanying text in Sec. 7.4, determine the minimum −3-dB beamwidth that will accommodate the tolerances in satellite position without the need for tracking.

7.9. Explain what is meant by *thermal control* and why this is necessary in a satellite.

7.10. Explain why an omnidirectional antenna must be used aboard a satellite for telemetry and command during the launch phase. How is the satellite powered during this phase?

7.11. Briefly describe the equipment sections making up a transponder channel.

7.12. Draw to scale the uplink and downlink channeling schemes for a 500-MHz-bandwidth C-band satellite, accommodating the full complement of 36-MHz-bandwidth transponders. Assume the use of 4-MHz guardbands.

7.13. Explain what is meant by *frequency reuse*, and describe briefly two methods by which this can be achieved.

7.14. Explain what is meant by *redundant receiver* in connection with communication satellites.

7.15. Describe the function of the input demultiplexer used aboard a communications satellite.

7.16. Describe briefly the most common type of high-power amplifying device used aboard a communications satellite.

7.17. What is the chief advantage of the TWTA used aboard satellites compared to other types of high-power amplifying devices? What are the main disadvantages of the TWTA?

7.18. Define and explain the term 1-dB *compression point*. What is the significance of this point in relation to the operating point of a TWTA?

7.19. Explain why operation near the saturation point of a TWTA is to be avoided when multiple carriers are being amplified simultaneously.

7.20. State the type of satellite antenna normally used to produce a widebeam radiation pattern, providing global coverage. How are spot beams produced?

7.21. Describe briefly how beam shaping of a satellite antenna radiation pattern may be achieved.

7.22. With reference to Figure 7.28, explain what is meant by a *four-for-two redundancy switching arrangement*.

References

CCIR. 1984. *Fixed Services Handbook*, final draft. Geneva.

Chetty, P. R. K. 1991. *Satellite Technology and Its Applications*. McGraw-Hill, New York.

Hyndman, J. E. 1991. *Hughes HS601 Communications Satellite Bus System Design Trades*. Hughes Aircraft Company, El Segundo, CA.

Johnston, E. C., and J. D. Thompson. 1982. "INTELSAT VI Communications Payload." *IEE Colloquium on the Global INTELSAT VI Satellite System*, Digest No. 1982/76. pp. 4/1–4/4.

Lilly, C. J. 1990. "INTELSAT's New Generation." *IEE Review*, Vol. 36, No. 3, March. pp. 111–113.

Pilcher, L. S. 1982. "Overall Design of the INTELSAT VI Satellite." *3rd International Conference on Satellite Systems for Mobile Communications and Navigation*, IEE, London.

Schwalb, A. 1982*a*. "The TIROS-N/NOAA-G Satellite Series." *NOAA Technical Memorandum NESS 95*, Washington, DC.

Schwalb, A. 1982*b*. "Modified Version of the TIROS-N/NOAA A-G Satellite Series (NOAA E-J): Advanced TIROS N (ATN)." *NOAA Technical Memorandum NESS 116*, Washington, DC.

Spilker, J. J. 1977. *Digital Communications by Satellite.* Prentice-Hall, Englewood Cliffs, NJ.

Wertz, J. R. (ed.). 1984. *Spacecraft Attitude Determination and Control.* D. Reidel, Holland.

Chapter

8

The Earth Segment

8.1 Introduction

The earth segment of a satellite communications system consists of the transmit and receive earth stations. The simplest of these are the home *TV receive-only* (TVRO) systems, and the most complex are the terminal stations used for international communications networks. Also included in the earth segment are those stations which are on ships at sea, and commercial and military land and aeronautical mobile stations.

As mentioned in Chap. 7, earth stations that are used for logistic support of satellites, such as providing the *telemetry, tracking, and command* (TT&C) functions, are considered as part of the space segment.

8.2 Receive-Only Home TV Systems

Planned broadcasting directly to home TV receivers takes place in the Ku (12-GHz) band. This service is known as *direct broadcast satellite* (DBS) service. There is some variation in the frequency bands assigned to different geographic regions. In the Americas, for example, the downlink band is 12.2 to 12.7 GHz, as described in Sec. 1.4.

The comparatively large satellite receiving dishes [ranging in diameter from about 1.83 m (6 ft) to about 3-m (10 ft) in some locations], which may be seen in some "backyards" are used to receive downlink TV signals at C band (4 GHz). Originally such downlink signals were never intended for home reception but for network relay to commercial TV outlets (VHF and UHF TV broadcast stations and cable TV "head-end" studios). Equipment is now marketed for home reception of C-band signals, and some manufacturers provide dual C-band/Ku-band equipment. A single mesh type reflector may be used which focuses the signals into a dual feed-horn, which has two separate outputs, one for the C-band signals and one

for the Ku-band signals. Much of television programming originates as *first generation signals*, also known as *master broadcast quality signals*. These are transmitted via satellite in the C band to the network head-end stations, where they are retransmitted as compressed digital signals to cable and direct broadcast satellite providers. One of the advantages claimed by sellers of C-band equipment for home reception is that there is no loss of quality compared with the compressed digital signals.

To take full advantage of C-band reception the home antenna has to be steerable to receive from different satellites, usually by means of a polar mount as described in Sec. 3.3. Another of the advantages, claimed for home C-band systems, is the larger number of satellites available for reception compared to what is available for direct broadcast satellite systems. Although many of the C-band transmissions are scrambled, there are free channels that can be received, and what are termed "wild feeds." These are also free, but unannounced programs, of which details can be found in advance from various publications and Internet sources. C-band users can also subscribe to pay TV channels, and another advantage claimed is that subscription services are cheaper than DBS or cable because of the multiple-source programming available.

The most widely advertised receiving system for C-band system appears to be 4DTV manufactured by Motorola. This enables reception of:

1. Free, analog signals and "wild feeds"
2. VideoCipher ll plus subscription services
3. Free DigiCipher 2 services
4. Subscription DigiCipher 2 services

VideoCipher is the brand name for the equipment used to scramble analog TV signals. DigiCipher 2 is the name given to the digital compression standard used in digital transmissions. General information about C-band TV reception will be found at http://orbitmagazine.com/ (Orbit, 2005) and http://www.satellitetheater.com/ (Satellite Theater systems, 2005).

The major differences between the Ku-band and the C-band receive-only systems lies in the frequency of operation of the outdoor unit and the fact that satellites intended for DBS have much higher *equivalent isotropic radiated power* (EIRP), as shown in Table 1.4. As already mentioned C-band antennas are considerably larger than DBS antennas. For clarity, only the Ku-band system is described here.

Figure 8.1 shows the main units in a home terminal DBS TV receiving system. Although there will be variations from system to system, the diagram covers the basic concept for analog [*frequency modulated* (FM)] TV. Direct-to-home digital TV, which is well on the way to replacing analog systems, is discussed in Chap. 16. However, the outdoor unit is similar for both systems.

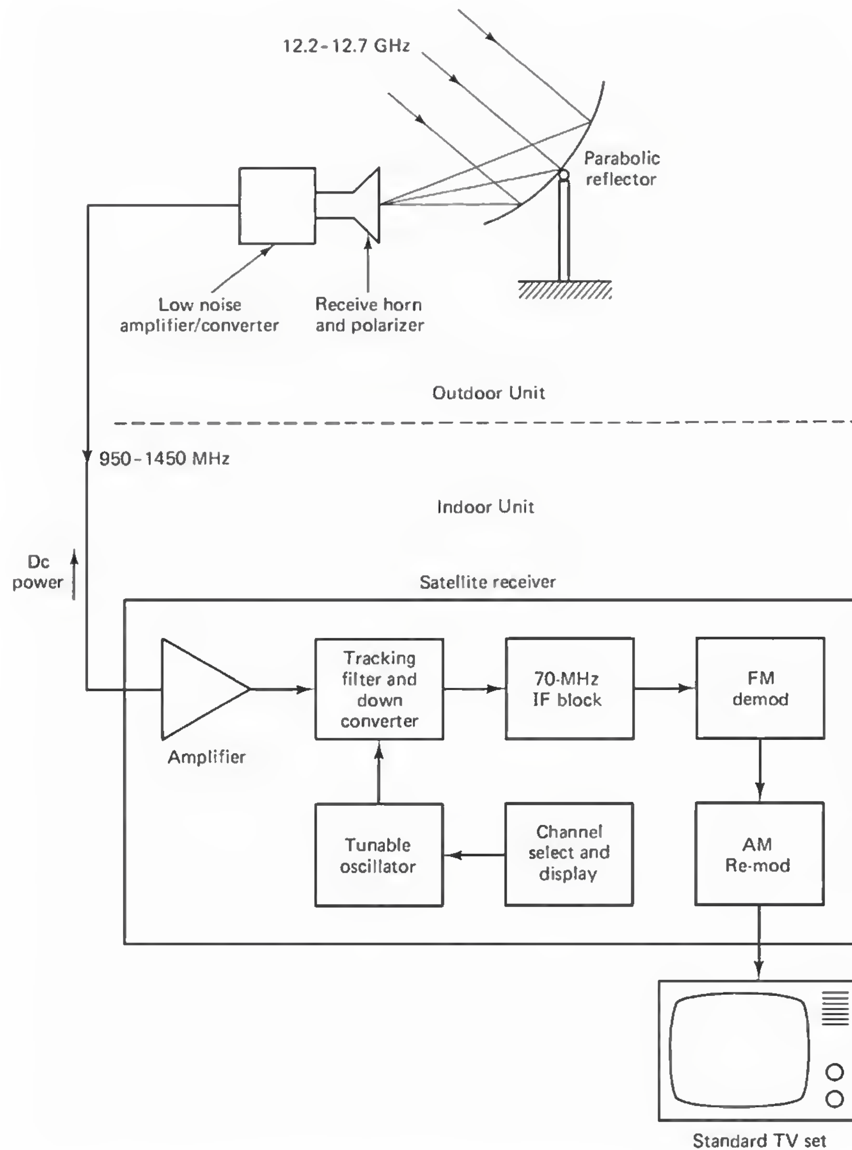


Figure 8.1 Block diagram showing a home terminal for DBS TV/FM reception.

8.2.1 The outdoor unit

This consists of a receiving antenna feeding directly into a low-noise amplifier/converter combination. A parabolic reflector is generally used, with the receiving horn mounted at the focus. A common design is to have the focus directly in front of the reflector, but for better interference rejection, an offset feed may be used as shown.

Huck and Day (1979) have shown that satisfactory reception can be achieved with reflector diameters in the range 0.6 to 1.6 m (1.97–5.25 ft), and the two nominal sizes often quoted are 0.9 m (2.95 ft) and 1.2 m (3.94 ft). By contrast, the reflector diameter for 4-GHz reception can range from 1.83 m (6 ft) to 3 m (10 ft). As noted in Sec. 6.13, the gain of a parabolic dish is proportional to $(D/\lambda)^2$. Comparing the gain of a 3-m dish at 4 GHz with a 1-m dish at 12 GHz, the ratio D/λ equals 40 in each case, so the gains will be about equal. Although the free-space losses are much higher at 12 GHz compared with 4 GHz, as described in Chap. 12, a higher-gain receiving antenna is not needed because the DBS operate at a much higher EIRP, as shown in Table 1.4.

The downlink frequency band of 12.2 to 12.7 GHz spans a range of 500 MHz, which accommodates 32 TV/FM channels, each of which is 24-MHz wide. Obviously, some overlap occurs between channels, but these are alternately polarized *left-hand circular* (LHC) and *right-hand circular* (RHC) or vertical/horizontal, to reduce interference to acceptable levels. This is referred to as *polarization interleaving*. A polarizer that may be switched to the desired polarization from the indoor control unit is required at the receiving horn.

The receiving horn feeds into a *low-noise converter* (LNC) or possibly a combination unit consisting of a *low-noise amplifier* (LNA) followed by a converter. The combination is referred to as an LNB, for *low-noise block*. The LNB provides gain for the broadband 12-GHz signal and then converts the signal to a lower frequency range so that a low-cost coaxial cable can be used as feeder to the indoor unit. The standard frequency range of this downconverted signal is 950 to 1450 MHz, as shown in Fig. 8.1. The coaxial cable, or an auxiliary wire pair, is used to carry dc power to the outdoor unit. Polarization-switching control wires are also required.

The low-noise amplification must be provided at the cable input in order to maintain a satisfactory signal-to-noise ratio. An LNA at the indoor end of the cable would be of little use, because it would also amplify the cable thermal noise. Single-to-noise ratio is discussed in more detail in Sec. 12.5. Of course, having to mount the LNB outside means that it must be able to operate over a wide range of climatic conditions, and homeowners may have to contend with the added problems of vandalism and theft.

8.2.2 The indoor unit for analog (FM) TV

The signal fed to the indoor unit is normally a wideband signal covering the range 950 to 1450 MHz. This is amplified and passed to a tracking filter which selects the desired channel, as shown in Fig. 8.1.

As previously mentioned, polarization interleaving is used, and only half the 32 channels will be present at the input of the indoor unit for any one setting of the antenna polarizer. This eases the job of the tracking filter, since alternate channels are well separated in frequency.

The selected channel is again downconverted, this time from the 950- to 1450-MHz range to a fixed intermediate frequency, usually 70 MHz although other values in the *very high frequency* (VHF) range are also used. The 70-MHz amplifier amplifies the signal up to the levels required for demodulation. A major difference between DBS TV and conventional TV is that with DBS, frequency modulation is used, whereas with conventional TV, amplitude modulation in the form of *vestigial single sideband* (VSSB) is used. The 70-MHz, FM *intermediate frequency* (IF) carrier therefore must be demodulated, and the baseband information used to generate a VSSB signal which is fed into one of the VHF/UHF channels of a standard TV set.

A DBS receiver provides a number of functions not shown on the simplified block diagram of Fig. 8.1. The demodulated video and audio signals are usually made available at output jacks. Also, as described in Sec. 13.3, an energy-dispersal waveform is applied to the satellite carrier to reduce interference, and this waveform has to be removed in the DBS receiver. Terminals also may be provided for the insertion of IF filters to reduce interference from terrestrial TV networks, and a descrambler also may be necessary for the reception of some programs. The indoor unit for digital TV is described in Chap. 16.

8.3 Master Antenna TV System

A *master antenna TV* (MATV) system is used to provide reception of DBS TV/FM channels to a small group of users, for example, to the tenants in an apartment building. It consists of a single outdoor unit (antenna and LNA/C) feeding a number of indoor units, as shown in Fig. 8.2. It is basically similar to the home system already described, but with each user having access to all the channels independently of the other users. The advantage is that only one outdoor unit is required, but as shown, separate LNA/Cs and feeder cables are required for each sense of polarization. Compared with the single-user system, a larger antenna is also required (2- to 3-m diameter) in order to maintain a good signal-to-noise ratio at all the indoor units.

Where more than a few subscribers are involved, the distribution system used is similar to the *community antenna* (CATV) system described in the following section.

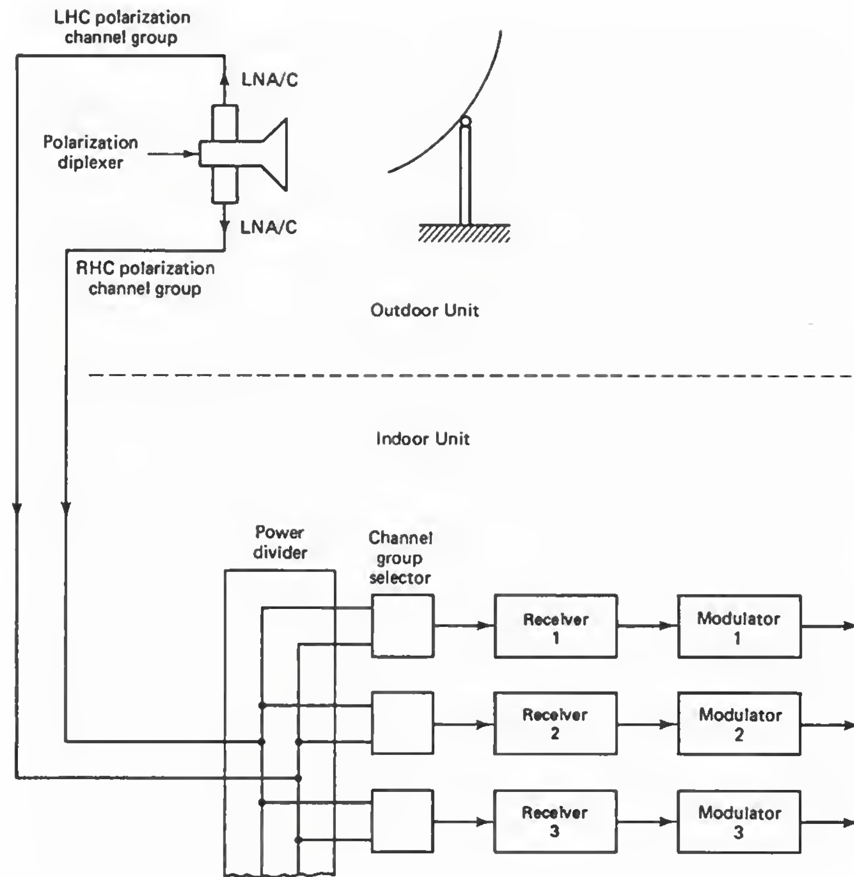


Figure 8.2 One possible arrangement for a master antenna TV (MATV) system.

8.4 Community Antenna TV System

The CATV system employs a single outdoor unit, with separate feeds available for each sense of polarization, like the MATV system, so that all channels are made available simultaneously at the indoor receiver. Instead of having a separate receiver for each user, all the carriers are demodulated in a common receiver-filter system, as shown in Fig. 8.3. The channels are then combined into a standard multiplexed signal for transmission over cable to the subscribers.

In remote areas where a cable distribution system may not be installed, the signal can be rebroadcast from a low-power VHF TV transmitter. Figure 8.4 shows a remote TV station which employs an

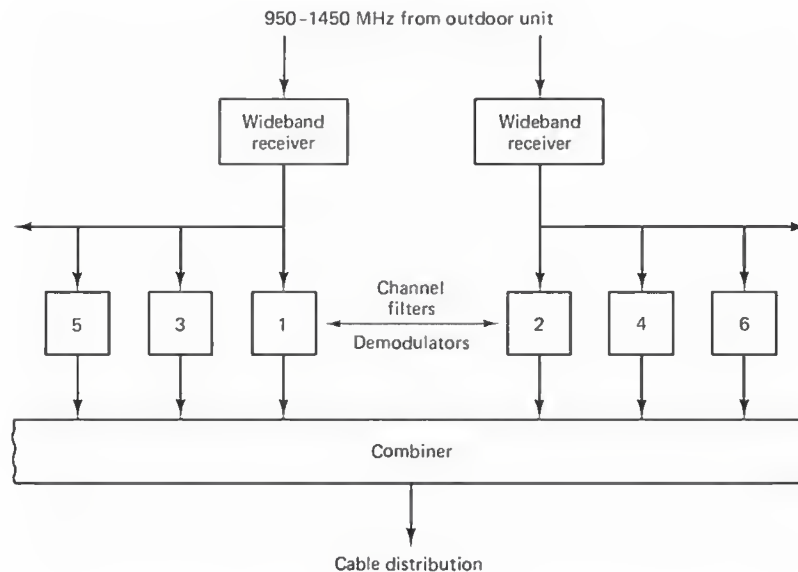


Figure 8.3 One possible arrangement for the indoor unit of a community antenna TV (CATV) system.

8-m (26.2-ft) antenna for reception of the satellite TV signal in the C band.

With the CATV system, local programming material also may be distributed to subscribers, an option which is not permitted in the MATV system.



Figure 8.4 Remote television station. (*Courtesy of Telesat Canada, 1983.*)

8.5 Transmit-Receive Earth Stations

In the previous sections, receive-only TV stations are described. Obviously, somewhere a transmit station must complete the uplink to the satellite. In some situations, a transmit-only station is required, for example, in relaying TV signals to the remote TVRO stations already described. Transmit-receive stations provide both functions and are required for telecommunications traffic generally, including network TV. The uplink facilities for digital TV are highly specialized and are covered in Chap. 16.

The basic elements for a redundant earth station are shown in Fig. 8.5. As mentioned in connection with transponders in Sec. 7.7.1, redundancy means that certain units are duplicated. A duplicate, or redundant, unit is automatically switched into a circuit to replace a corresponding unit that has failed. Redundant units are shown by dashed lines in Fig. 8.5.

The block diagram is shown in more detail in Fig. 8.6, where, for clarity, redundant units are not shown. Starting at the bottom of the diagram, the first block shows the interconnection equipment required between satellite station and the terrestrial network. For the purpose of explanation, telephone traffic will be assumed. This may consist of a number of telephone channels in a multiplexed format. Multiplexing is a method of grouping telephone channels together, usually in basic groups of 12, without mutual interference. It is described in detail in Chaps. 9 and 10.

It may be that groupings different from those used in the terrestrial network are required for satellite transmission, and the next block shows the multiplexing equipment in which the reformatting is carried out. Following along the transmit chain, the multiplexed signal is modulated onto a carrier wave at an intermediate frequency, usually 70 MHz. Parallel IF stages are required, one for each microwave carrier to be transmitted. After amplification at the 70-MHz IF, the modulated signal is then upconverted to the required microwave carrier frequency. A number of carriers may be transmitted simultaneously, and although these are at different frequencies they are generally specified by their nominal frequency, for example, as 6-GHz or 14-GHz carriers.

It should be noted that the individual carriers may be multidestination carriers. This means that they carry traffic destined for different stations. For example, as part of its load, a microwave carrier may have telephone traffic for Boston and New York. The same carrier is received at both places, and the designated traffic sorted out by filters at the receiving earth station.

Referring again to the block diagram of Fig. 8.6, after passing through the upconverters, the carriers are combined, and the resulting wideband signal is amplified. The wideband power signal is fed to the antenna

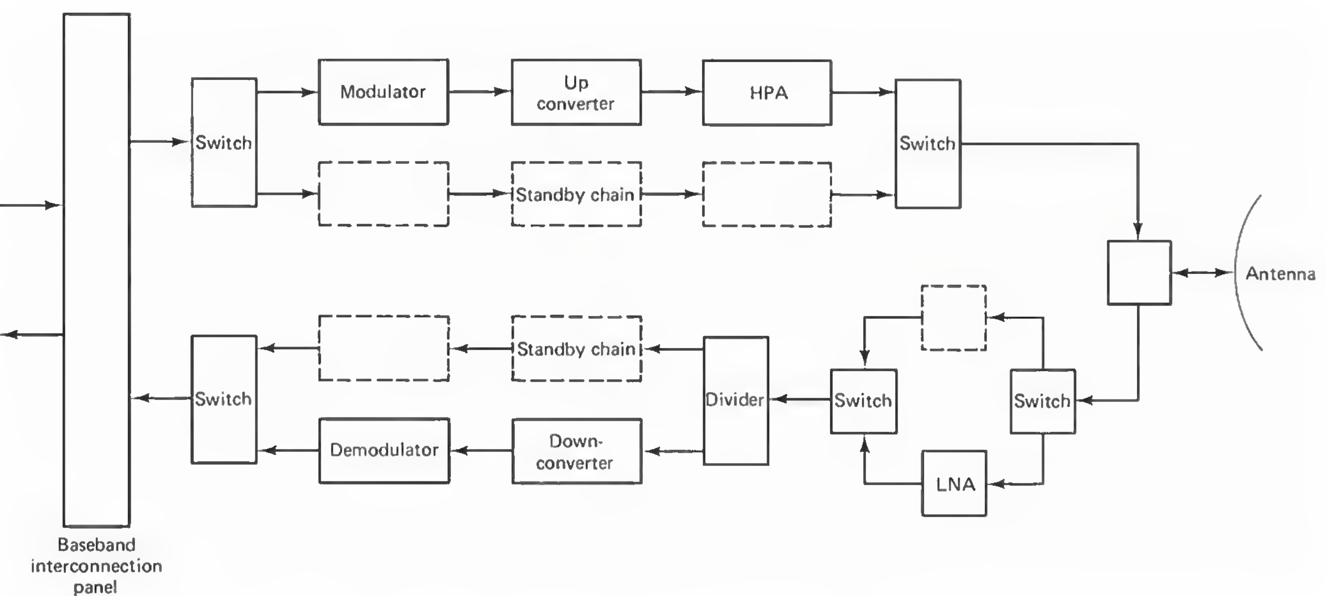


Figure 8.5 Basic elements of a redundant earth station. (*Courtesy of Telesat Canada, 1983.*)

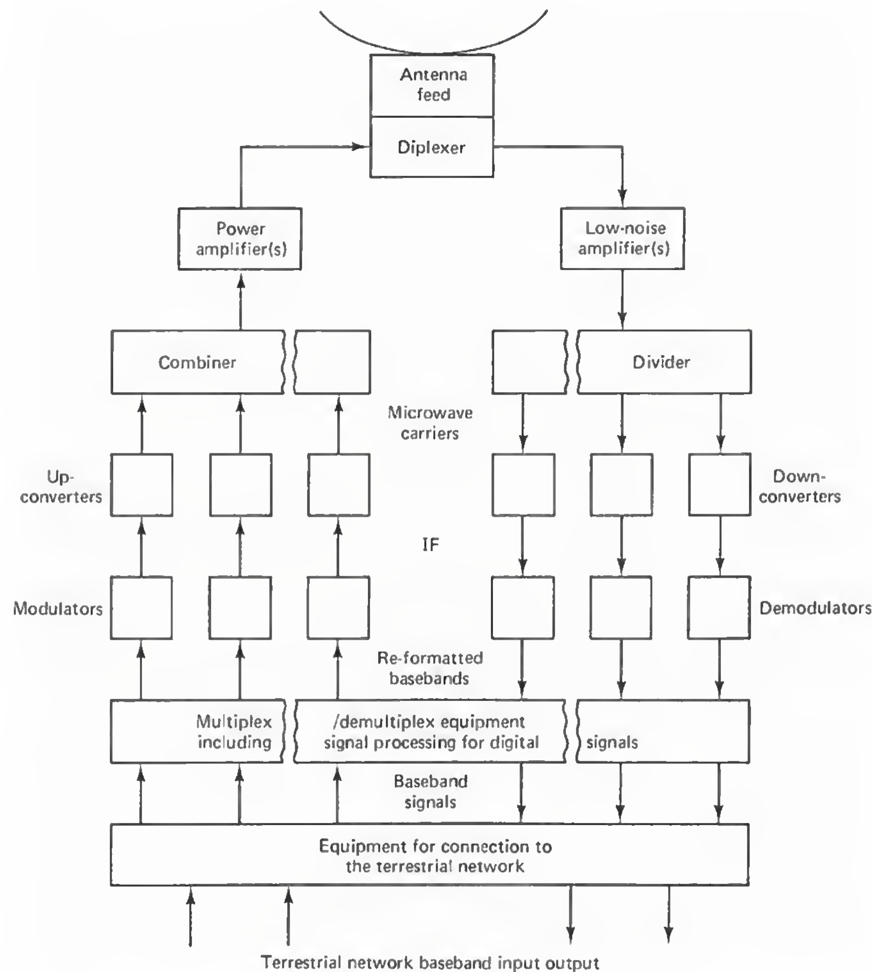


Figure 8.6 More detailed block diagram of a transmit-receive earth station.

through a diplexer, which allows the antenna to handle transmit and receive signals simultaneously.

The station's antenna functions in both, the transmit and receive modes, but at different frequencies. In the C band, the nominal uplink, or transmit, frequency is 6 GHz and the downlink, or receive, frequency is nominally 4 GHz. In the Ku band, the uplink frequency is nominally 14 GHz, and the downlink, 12 GHz. High-gain antennas are employed in both bands, which also means narrow antenna beams. A narrow beam is necessary to prevent interference between neighboring satellite links. In the case of C band, interference to and from terrestrial microwave

links also must be avoided. Terrestrial microwave links do not operate at Ku-band frequencies.

In the receive branch (the right-hand side of Fig. 8.6), the incoming wideband signal is amplified in an LNA and passed to a divider network, which separates out the individual microwave carriers. These are each downconverted to an IF band and passed on to the multiplex block, where the multiplexed signals are reformatted as required by the terrestrial network.

It should be noted that, in general, the signal traffic flow on the receive side will differ from that on the transmit side. The incoming microwave carriers will be different in number and in the amount of traffic carried, and the multiplexed output will carry telephone circuits not necessarily carried on the transmit side.

A number of different classes of earth stations are available, depending on the service requirements. Traffic can be broadly classified as heavy route, medium route, and thin route. In a thin-route circuit, a transponder channel (36 MHz) may be occupied by a number of single carriers, each associated with its own voice circuit. This mode of operation is known as *single carrier per channel* (SCPC), a multiple-access mode which is discussed further in Chap. 14. Antenna sizes range from 3.6 m (11.8 ft) for transportable stations up to 30 m (98.4 ft) for a main terminal.

A medium-route circuit also provides multiple access, either on the basis of *frequency-division multiple access* (FDMA) or *time-division multiple access* (TDMA), multiplexed baseband signals being carried in either case. These access modes are also described in detail in Chap. 14. Antenna sizes range from 30 m (89.4 ft) for a main station to 10 m (32.8 ft) for a remote station.

In a 6/4-GHz heavy-route system, each satellite channel (bandwidth 36 MHz) is capable of carrying over 960 one-way voice circuits simultaneously or a single-color analog TV signal with associated audio (in some systems two analog TV signals can be accommodated). Thus the transponder channel for a heavy-route circuit carries one large-bandwidth signal, which may be TV or multiplexed telephony. The antenna diameter for a heavy-route circuit is at least 30 m (98.4 ft). For international operation such antennas are designed to the INTELSAT specifications for a Standard A earth station (Intelsat, 1982). Figure 8.7 shows a photograph of a 32-m (105-ft) Standard A earth station antenna.

It will be appreciated that for these large antennas, which may weigh in the order of 250 tons, the foundations must be very strong and stable. Such large diameters automatically mean very narrow beams, and therefore, any movement which would deflect the beam unduly must be avoided. Where snow and ice conditions are likely to be encountered, built-in heaters are required. For the antenna shown in Fig. 8.7, deicing



Figure 8.7 Standard-A (C-band 6/4 GHz) 32-m antenna. (*Courtesy of TIW Systems, Inc., Sunnydale, CA.*)

heaters provide reflector surface heat of $40 W/ft^2$ for the main reflectors and subreflectors, and 3000 W for the azimuth wheels.

Although these antennas are used with geostationary satellites, some drift in the satellite position does occur, as shown in Chap. 3. This, combined with the very narrow beams of the larger earth station antennas, means that some provision must be made for a limited degree of tracking. Step adjustments in azimuth and elevation may be made, under computer control, to maximize the received signal.

The continuity of the primary power supply is another important consideration in the design of transmit-receive earth stations. Apart from the smallest stations, power backup in the form of multiple feeds from the commercial power source and/or batteries and generators is provided. If the commercial power fails, batteries immediately take over with no interruption. At the same time, the standby generators start up, and once they are up to speed they automatically take over from the batteries.

8.6 Problems and Exercises

8.1. Explain what is meant by DBS service. How does this differ from the home reception of satellite TV signals in the C band?

8.2. Explain what is meant by *polarization interleaving*. On a frequency axis, draw to scale the channel allocations for the 32 TV channels in the Ku band, showing how polarization interleaving is used in this.

8.3. Why is it desirable to downconvert the satellite TV signal received at the antenna?

8.4. Explain why the LNA in a satellite receiving system is placed at the antenna end of the feeder cable.

8.5. With the aid of a block schematic, briefly describe the functioning of the indoor receiving unit of a satellite TV/FM receiving system intended for home reception.

8.6. In most satellite TV receivers the first IF band is converted to a second, fixed IF. Why is this second frequency conversion required?

8.7. For the standard home television set to function in a satellite TV/FM receiving system, a demodulator/remodulator unit is needed. Explain why.

8.8. Describe and compare the MATV and the CATV systems.

8.9. Explain what is meant by the term *redundant earth station*.

8.10. With the aid of a block schematic, describe the functioning of a transmit-receive earth station used for telephone traffic. Describe a multidestination carrier.

References

Huck, R. W., and J. W. B. Day. 1979. "Experience in Satellite Broadcasting Applications with CTS/HERMES." *XIth International TV Symposium*, Montreux, 27 May–1 June.

INTELSAT. 1982. "Standard A Performance Characteristics of Earth Stations in the INTELSAT IV, IVA, and V Systems." BG-28-72E M/6/77.

Orbit, 2005, at http://orbitmagazine.com/

Satellite Theater systems, 2005, at http://www.satellitetheater.com/

Chapter

9

Analog Signals

9.1 Introduction

Analog signals are electrical replicas of the original signals such as audio and video. *Baseband signals* are those signals which occupy the lowest, or base, band of frequencies, in the frequency spectrum used by the telecommunications network. A baseband signal may consist of one or more information signals. For example, a number of analog telephony signals may be combined into one baseband signal by the process known as *frequency-division multiplexing* (FDM). Other common types of baseband signals are the multiplexed video and audio signals which originate in the TV studio. In forming the multiplexed baseband signals, the information signals are *modulated* onto subcarriers. This modulation step must be distinguished from the modulation process, which places the multiplexed signal onto the microwave carrier for transmission to the satellite.

In this chapter, the characteristics of the more common types of analog baseband signals are described, along with representative methods of analog modulation.

9.2 The Telephone Channel

Natural speech, including that of female and male voices, covers a frequency range of about 80 to 8000 Hz. The somewhat unnatural quality associated with telephone speech results from the fact that a considerably smaller band of frequencies is used for normal telephone transmission. The range of 300 to 3400 Hz is accepted internationally as the standard for "telephone quality" speech, and this is termed the *speech baseband*. In practice, some variations occur in the basebands used by different telephone companies. The telephone channel is often referred

to as a *voice frequency* (VF) channel, and in this book this will be taken to mean the frequency range of 300 to 3400 Hz.

There are good reasons for limiting the frequency range. Noise, which covers a very wide frequency spectrum, is reduced by reducing the bandwidth. Also, reducing the bandwidth allows more telephone channels to be carried over a given type of circuit, as will be described in Sec. 9.4.

The signal levels encountered within telephone networks vary considerably. Audio signal levels are often measured in *volume units* (VU). For a sinusoidal signal within the VF range, 0 VU corresponds to 1 mW of power, or 0 dBm. No simple relationship exists between VU and power for speech signals, but as a rough guide, the power level in dBm of normal speech is given by VU −1.4. As a rule of thumb, the average voice level on a telephone circuit (or mean talker level) is defined as −13 VU (see Freeman, 1981).

9.3 Single-Sideband Telephony

Figure 9.1*a* shows how the VF baseband may be represented in the frequency domain. In some cases, the triangular representation has the small end of the triangle at 0 Hz, even though frequency components below 300 Hz actually may not be present. Also, in some cases, the upper end is set at 4 kHz to indicate allowance for a guard band, the need for which will be described later.

When the telephone signal is multiplied in the time domain with a sinusoidal carrier of frequency *fc*, a new spectrum results, in which the original baseband appears on either side of the carrier frequency. This is illustrated in Fig. 9.1*b* for a carrier of 20 kHz, where the band of frequencies below the carrier is referred to as the *lower sideband* and the band above the carrier as the *upper sideband*. To avoid distortion which would occur with sideband overlap, the carrier frequency must be greater than the highest frequency in the baseband.

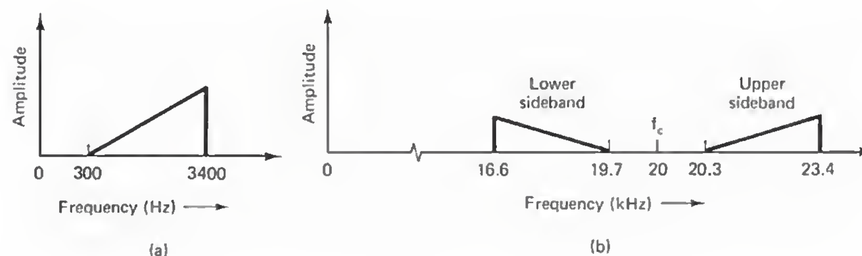


Figure 9.1 Frequency-domain representation of (*a*) a telephone baseband signal and (*b*) the double-sideband suppressed carrier (DSBSC) modulated version of (*a*).

The result of this multiplication process is referred to as *double-sideband suppressed-carrier* (DSBSC) modulation, since only the sidebands, and not the carrier, appear in the spectrum. Now, all the information in the original telephone signal is contained in either of the two sidebands, and therefore, it is necessary to transmit only one of these. A filter may be used to select either one and reject the other. The resulting output is termed a *single-sideband* (SSB) signal.

The SSB process utilizing the lower sideband is illustrated in Fig. 9.2, where a 20-kHz carrier is used as an example. It will be seen that for the lower sideband, the frequencies have been inverted, the highest baseband frequency being translated to the lowest transmission frequency at 16.6 kHz and the lowest baseband frequency to the highest transmission frequency at 19.7 kHz. This inversion does not affect the final baseband output, since the demodulation process reinverts the spectrum. At the receiver, the SSB signal is demodulated (i.e., the baseband signal is recovered) by being passed through an identical multiplying process. In this case the multiplying sinusoid, termed the *local oscillator* (LO) signal, must have the same frequency as the original carrier. A low-pass filter is required at the output to select the baseband signal and reject other, unwanted frequency components generated in the demodulation process. This single-sideband modulation/demodulation process is illustrated in Fig. 9.2.

The way in which SSB signals are used for the simultaneous transmission of a number of telephone signals is described in the following section. It should be noted at this point that a number of different carriers are likely to be used in a satellite link. The *radiofrequency* (rf) carrier used in transmission to and from the satellite will be much higher in frequency than those used for the generation of the set of SSB signals. These latter carriers are sometimes referred to as VF *carriers*. The term

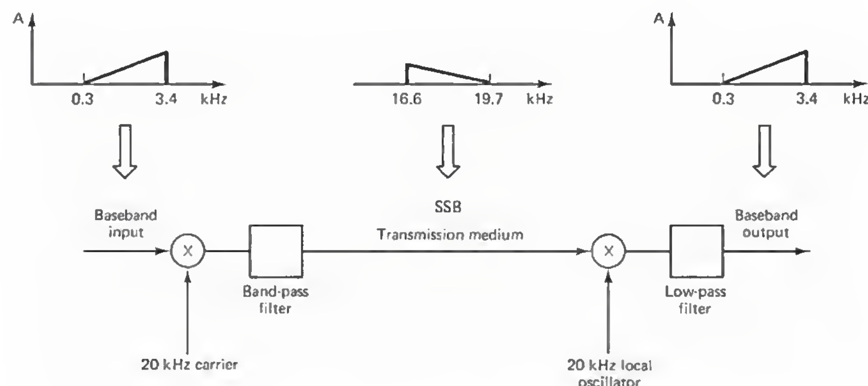


Figure 9.2 A basic SSB transmission scheme.

subcarrier is also used, and this practice will be followed here. Thus the 20-kHz carrier shown in Fig. 9.2 is a subcarrier.

Companded single sideband (CSSB) refers to a technique in which the speech signal levels are compressed before transmission as a single sideband, and at the receiver they are expanded again back to their original levels. (The term *compander* is derived from *comp*ressor-exp*ander*). In one companded system described by Campanella (1983), a 2:1 compression in decibels is used, followed by a 1:2 expansion at the receiver. It is shown in the reference that the expander decreases its attenuation when a speech signal is present and increases its attenuation when it is absent. In this way the "idle" noise on the channel is reduced, which allows the channel to operate at a reduced carrier-to-noise ratio. This in turn permits more channels to occupy a given satellite link, a topic which comes under the heading of *multiple access* and which is described more fully in Chap. 14.

9.4 FDM Telephony

FDM provides a way of keeping a number of individual telephone signals separate while transmitting them simultaneously over a common transmission link circuit. Each telephone baseband signal is modulated onto a separate subcarrier, and all the upper or all the lower sidebands are combined to form the frequency-multiplexed signal. Figure 9.3*a* shows how three voice channels may be frequency-division multiplexed. Each voice channel occupies the range 300 to 3400 Hz, and each is modulated onto its own subcarrier. The subcarrier frequency separation is 4 kHz, allowing for the basic voice bandwidth of 3.1 kHz plus an adequate guardband for filtering. The upper sidebands are selected by means of filters and then combined to form the final three-channel multiplexed signal. The three-channel FDM *pregroup* signal can be represented by a single triangle, as shown in Fig. 9.3*b*.

To facilitate interconnection among the different telecommunications systems in use worldwide, the *Comité Consultatif Internationale de Télégraphique et Téléphonique* (CCITT)* has recommended a standard modulation plan for FDM (CCITT G322 and G423, 1976). The standard group in the plan consists of 12 voice channels. One way to create such groups is to use an arrangement similar to that shown in Fig. 9.3*a*, except of course that 12 multipliers and 12 sideband filters are required. In the standard plan, the lower sidebands are selected by the filters, and the *group* bandwidth extends from 60 to 108 kHz.

As an alternative to forming a 12-channel group directly, the VF channels may be frequency-division multiplexed in threes by using the

*Since 1994, the CCITT has been reorganized by the International Telecommunications Union (ITU) into a new sector ITU-T.

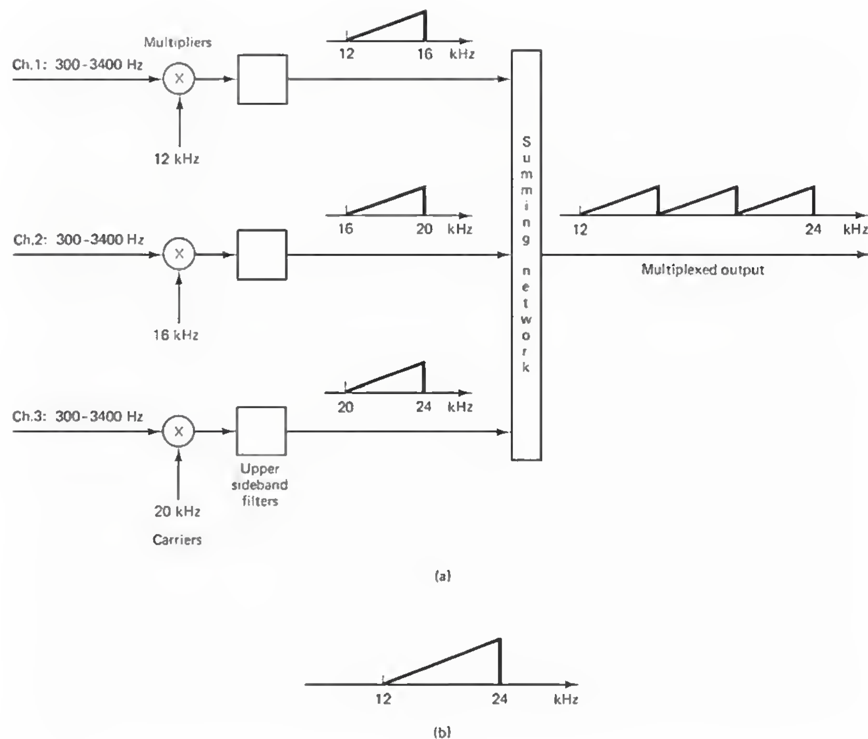


Figure 9.3 (*a*) Three-channel frequency-division multiplex scheme; (*b*) simplified representation.

arrangement shown in Fig. 9.3*a*. The four 3-channel-multiplexed signals, termed pregroups, are then combined to form the 12-channel group. This approach eases the filtering requirements but does require an additional mixer stage, which adds noise to the process.

The main group designations in the CCITT modulation plan are:

Group. As already described, this consists of 12 VF channels, each occupying a 4-kHz bandwidth in the multiplexed output. The overall bandwidth of a group extends from 60 to 108 kHz.

Supergroup. A supergroup is formed by FDM five groups together. The lower sidebands are combined to form a 60-VF-channel supergroup extending from 312 to 552 kHz.

Basic mastergroup. A basic mastergroup is formed by FDM five supergroups together. The lower sidebands are combined to form a 300-VF-channel basic mastergroup.

Allowing for 8-kHz guard bands between sidebands, the basic mastergroup extends from 812 to 2044 kHz.

Super mastergroup. A super mastergroup is formed by FDM three basic mastergroups together. The lower sidebands are combined to form the 900-VF-channel super mastergroup. Allowing for 8-kHz guardbands between sidebands, the super mastergroup extends from 8516 to 12,388 kHz.

In satellite communications, such multiplexed signals often form the baseband signal which is used to frequency modulate a microwave carrier (see Sec. 9.6). The smallest baseband unit is usually the 12-channel VF group, and larger groupings are multiples of this unit. Figure 9.4 shows how 24-, 60-, and 252-VF-channel baseband signals may be formed. These examples are taken from CCITT Recommendations G322 and G423. It will be observed that in each case a group occupies the range 12 to 60 kHz. Because of this, the 60-VF-channel baseband, which modulates the carrier to the satellite, differs somewhat from the standard 60-VF-channel supergroup signal used for terrestrial cable or microwave FDM links.

9.5 Color Television

The baseband signal for television is a composite of the visual information signals and synchronization signals. The visual information is transmitted as three signal components, denoted as the Y, I, and Q signals. The Y signal is a *luminance*, or *intensity*, component and is also the only visual information signal required by monochrome receivers. The I and Q signals are termed *chrominance components*, and together they convey information on the hue or tint and on the amount of saturation of the coloring which is present.

The synchronization signal consists of narrow pulses at the end of each line scan for horizontal synchronization and a sequence of narrower and wider pulses at the end of each field scan for vertical synchronization. Additional synchronization for the color information demodulation in the receiver is superimposed on the horizontal pulses, as described below.

The luminance signal and the synchronization pulses require a base bandwidth of 4.2 MHz for North American standards. The baseband extends down to and includes a dc component. The composite signal containing the luminance and synchronization information is designed to be fully compatible with the requirements of monochrome (black-and-white) receivers.

In transmitting the chrominance information, use is made of the fact that the eye cannot resolve colors to the extent that it can resolve intensity detail; this allows the chrominance signal bandwidth to be less than that of the luminance signal. The I and Q chrominance signals are

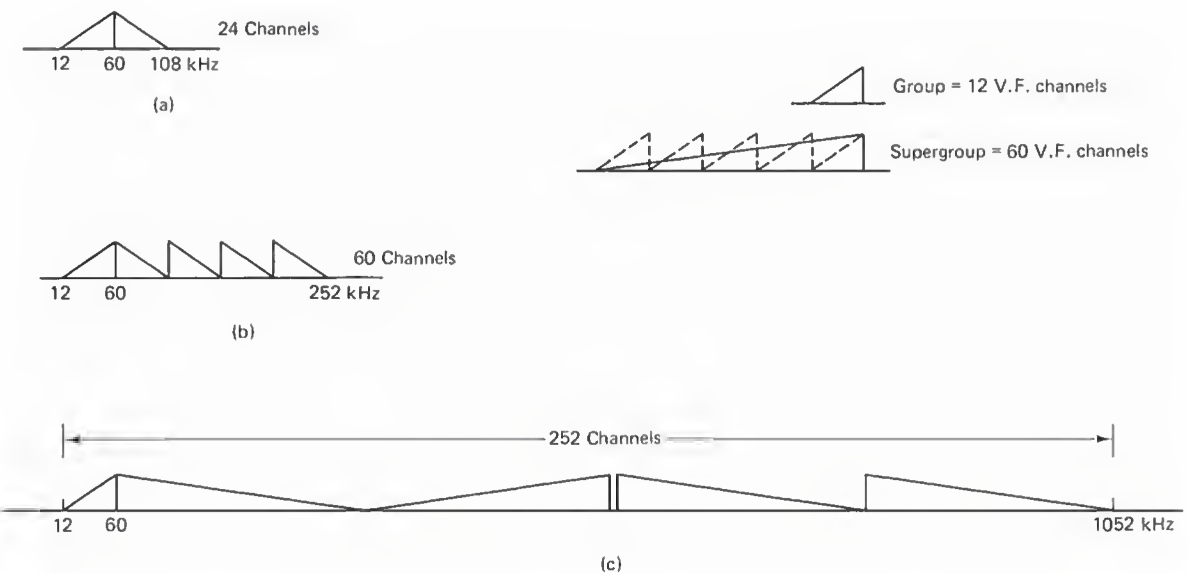


Figure 9.4 Examples of baseband signals for FDM telephony: (*a*) 24 channels; (*b*) 60 channels; (*c*) 252 channels.

transmitted within the luminance bandwidth by quadrature DSBSC (as seen later), modulating them onto a subcarrier which places them at the upper end of the luminance signal spectrum. Use is made of the fact that the eye cannot readily perceive the interference which results when the chrominance signals are transmitted within the luminance signal bandwidth. The baseband response is shown in Fig. 9.5.

Different methods of chrominance subcarrier modulation are employed in different countries. In France, a system known as *sequential couleur a mémoire* (SECAM) is used. In most other European countries, a system known as *phase alternation line* (PAL) is used. In North America, the NTSC system is used, where NTSC stands for *National Television System Committee.*

In the NTSC system, each chrominance signal is modulated onto its subcarrier using DSBSC modulation, as described in Sec. 9.3. A single oscillator source is used so that the I and Q signal subcarriers have the same frequency, but one of the subcarriers is shifted 90° in phase to preserve the separate chrominance information in the I and Q baseband signals. This method is known as *quadrature modulation* (QM). The I signal is the chrominance signal which modulates the in-phase carrier. Its bandwidth in the NTSC system is restricted to 1.5 MHz, and after modulation onto the subcarrier, a single-sideband filter removes the upper sideband components more than 0.5 MHz above the carrier. This is referred to as a *vestigial sideband* (VSB). The modulated I signal therefore consists of the 1.5-MHz lower sideband plus the 0.5 MHz upper VSB.

The Q signal is the chrominance signal which modulates the quadrature carrier. Its bandwidth is restricted to 0.5 MHz, and after modulation, a DSBSC signal results. The spectrum magnitude of the combined I and Q signals is shown in Fig. 9.5.

The magnitude of the QM envelope contains the color saturation information, and the phase angle of the QM envelope contains the hue, or tint,

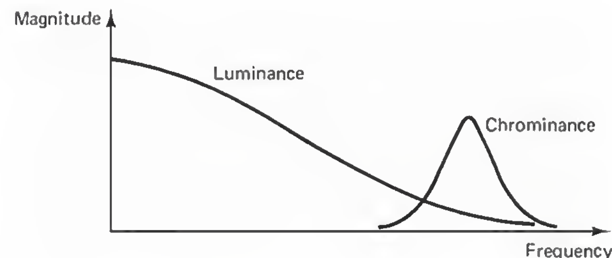


Figure 9.5 Frequency spectra for the luminance and chrominance signals.

information. The chrominance signal subcarrier frequency has to be precisely controlled, and in the NTSC system it is held at 3.579545 MHz ± 10 Hz, which places the subcarrier frequency midway between the 227th and the 228th harmonics of the horizontal scanning rate (frequency). The luminance and chrominance signals are both characterized by spectra wherein the power spectral density occurs in groups which are centered about the harmonics of the horizontal scan frequency. Placing the chrominance subcarrier midway between the 227th and 228th horizontal-scan harmonics of the luminance-plus-synchronization signals causes the luminance and the chrominance signals to be interleaved in the spectrum of the composite NTSC signal. This interleaving is most apparent in the range from about 3.0 to 4.1 MHz. The presence of the chrominance signal causes high-frequency modulation of the luminance signal and produces a very fine stationary dot-matrix pattern in the picture areas of high color saturation. To prevent this, most of the cheaper TV receivers limit the luminance channel video bandwidth to about 2.8 to 3.1 MHz. More expensive "high resolution" receivers employ a *comb filter* to remove most of the chrominance signal from the luminance-channel signal while still maintaining about a 4-MHz luminance-channel video bandwidth.

Because the subcarrier is suppressed in the modulation process, a subcarrier frequency and phase reference carrier must be transmitted to allow the I and Q baseband chrominance signals to be demodulated at the receiver. This reference signal is transmitted in the form of bursts of 8 to 11 cycles of the phase-shifted subcarrier, transmitted on the "backporch" of the horizontal blanking pulse. These bursts are transmitted toward the end of each line sync period, part of the line sync pulse being suppressed to accommodate them. One line waveform including the synchronization signals is shown in Fig. 9.6.

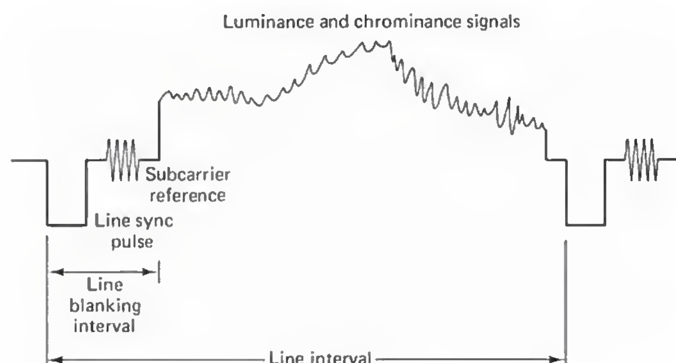


Figure 9.6 One line of waveform for a color TV signal.

Figure 9.7 shows in block schematic form the NTSC system. The TV camera contains three separate camera tubes, one for each of the colors red, blue, and green. It is known that colored light can be synthesized by *additive* mixing of red, blue, and green light beams, these being the three primary light beam colors. For example, yellow is obtained by adding red and green light. (This process must be distinguished from the subtractive process of paint pigments, in which the primary pigment colors are red, blue, and yellow.)

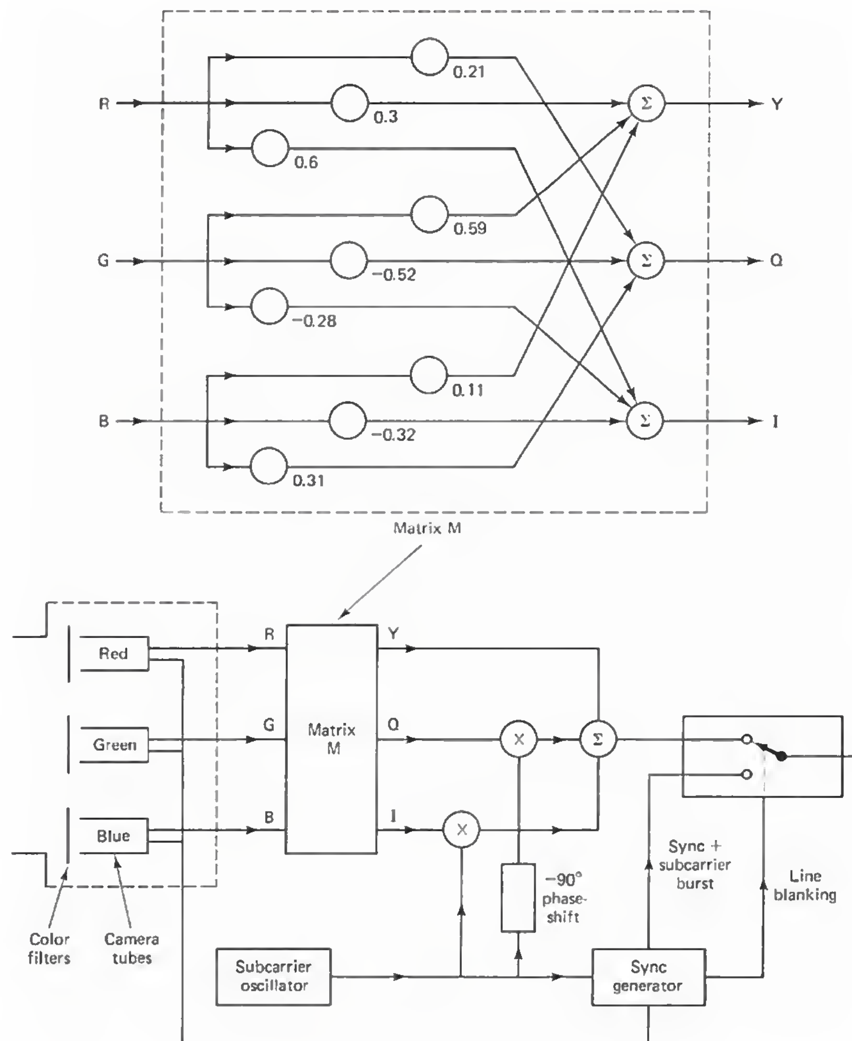


Figure 9.7 Generation of NTSC color TV signal. Matrix M converts the three color signals into the luminance and chrominance signals.

Color filters are used in front of each tube to sharpen its response. In principle, it would be possible to transmit the three color signals and at the receiver reconstruct the color scene from them. However, this is not the best technical approach because such signals would not be compatible with monochrome television and would require extra bandwidth. Instead, three new signals are generated which do provide compatibility and do not require extra bandwidth. These are the luminance signal and the two chrominance signals which have been described already. The process of generating the new signals from the color signals is mathematically equivalent to having three equations in three variables and rearranging these in terms of three new variables which are linear combinations of the original three. The details are shown in the matrix M block of Fig. 9.7, and derivation of the equations from this is left as Prob. 9.9.

At the receiver, the three color signals can be synthesized from the luminance and chrominance components. Again, this is mathematically equivalent to rearranging the three equations into their original form. The three color signals then modulate the electron beams which excite the corresponding color phosphors in the TV tube. The complete video signal is therefore a multiplexed baseband signal which extends from dc up to 4.2 MHz and which contains all the visual information plus synchronization signals.

In conventional TV broadcasting, the aural signal is transmitted by a separate transmitter, as shown in Fig. 9.8*a*. The aural information is received by stereo microphones, split into (L + R) and (L − R) signals, where L stands for left and R for right. The (L − R) signal is used to DSBSC modulate a subcarrier at $2f_h$ (31.468 kHz). This DSBSC signal is then added to the (L + R) signal and used to frequency modulate a separate transmitter whose rf carrier frequency is 4.5 MHz above the rf carrier frequency of the video transmitter. The outputs of these two transmitters may go to separate antennas or may be combined and fed into a single antenna, as is shown in Fig. 9.8*a*.

The signal format for satellite analog TV differs from that of conventional TV, as shown in Fig. 9.8. To generate the uplink microwave TV signal to a communications satellite transponder channel, the composite video signal (going from 0 Hz to about 4.2 MHz for the North American NTSC standard) is added to two or three *frequency modulation* (FM) carriers at frequencies of 6.2, 6.8, and/or 7.4 MHz, which carry audio information. This composite FDM signal is then, in turn, used to frequency modulate the uplink microwave carrier signal, producing a signal with an rf bandwidth of about 36 MHz. The availability of three possible audio signal carriers permits the transmission of stereo and/or multilingual audio over the satellite link. Figure 9.8*b* shows a block diagram of this system.

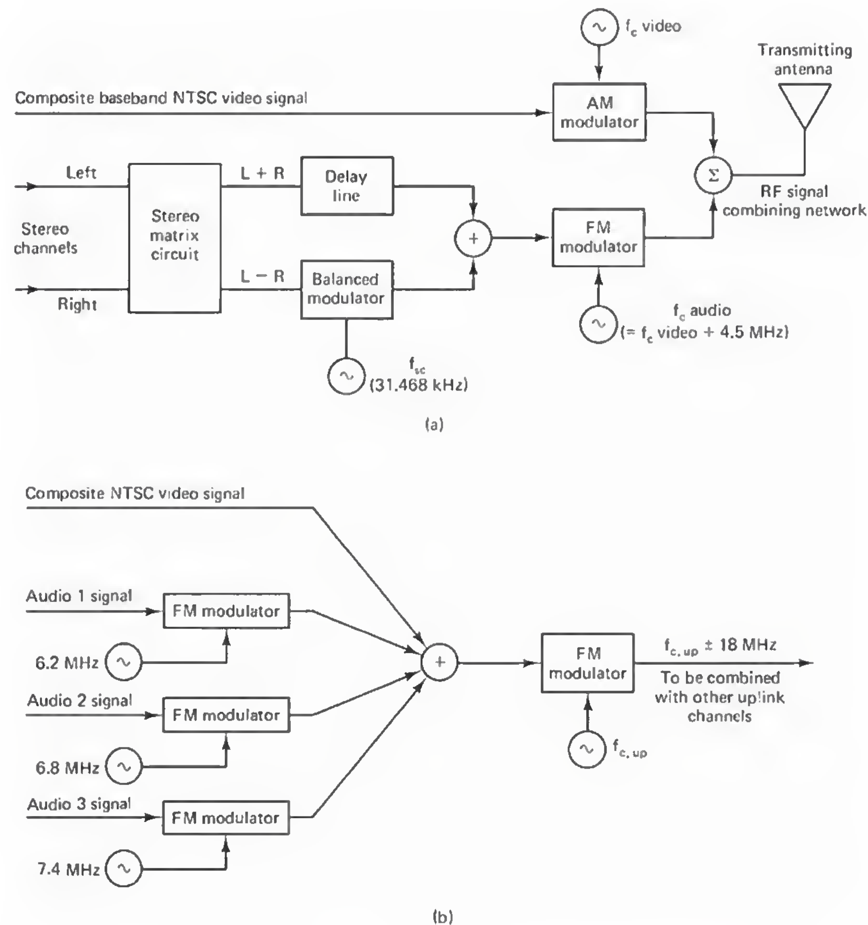


Figure 9.8 (*a*) Conventional analog TV broadcasting of the video and aural signals; (*b*) generation of a satellite uplink signal for analog TV.

As mentioned previously, three color TV systems—NTSC, PAL, and SECAM—are in widespread use. In addition, different countries use different line frequencies (determined by the frequency of the domestic power supply) and different numbers of lines per scan. Broadcasting between countries utilizing different standards requires the use of a converter. Transmission takes place using the standards of the country originating the broadcast, and conversion to the standards of the receiving country takes place at the receiving station. The conversion may take place through optical image processing or by conversion of the electronic signal format. The latter can be further subdivided into analog and digital techniques. The digital converter, referred to as

digital intercontinental conversion equipment (DICE), is favored because of its good performance and lower cost (see Miya, 1981).

9.6 Frequency Modulation

The analog signals discussed in the previous sections are transferred to the microwave carrier by means of FM. Instead of being done in one step, as shown in Fig. 9.8*b*, this modulation usually takes place at an intermediate frequency, as shown in Fig. 8.6. This signal is then frequency multiplied up to the required uplink microwave frequency. In the receive branch of Fig. 8.6, the incoming (downlink) FM microwave signal is downconverted to an intermediate frequency, and the baseband signal is recovered from the *intermediate frequency* (IF) carrier in the demodulator. The actual baseband video signal is now available directly via a low-pass filter, but the audio channels must each undergo an additional step of FM demodulation to recover the baseband audio signals.

A major advantage associated with FM is the improvement in the postdetection signal-to-noise ratio at the receiver output compared with other analog modulation methods. This improvement can be attributed to three factors:

1. Amplitude limiting
2. A property of FM which allows an exchange between signal-to-noise ratio and bandwidth
3. A noise reduction inherent in the way noise phase modulates a carrier

These factors are discussed in more detail in the following sections.

Figure 9.9 shows the basic circuit blocks of an FM receiver. The receiver noise, including that from the antenna, can be lumped into one equivalent noise source at the receiver input, as described in Sec. 12.5. It is emphasized at this point that thermal-like noise only is being considered, the main characteristic of which is that the spectral density of the noise power is constant, as given by Eq. (12.15). This is referred to as a *flat spectrum*. (This type of noise is also referred to as *white noise* in analogy to white light, which contains a uniform spectrum of colors.) Both the signal spectrum and the noise spectrum are converted to the

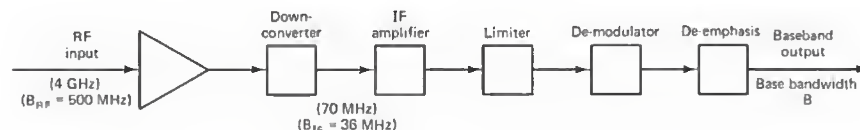


Figure 9.9 Elements of an FM receiver. Figures shown in parentheses are typical.

intermediate frequency bands, with the bandwidth of the IF stage determining the total noise power at the input to the demodulator. The IF bandwidth has to be wide enough to accommodate the FM signal, as described in Sec. 9.6.2, but should be no wider.

9.6.1 Limiters

The total thermal noise referred to the receiver input modulates the incoming carrier in amplitude and in phase. The rf limiter circuit (often referred to as an instantaneous or "hard" limiter) following the IF amplifier removes the amplitude modulation, leaving only the phase-modulation component of the noise. The limiter is an amplifier designed to operate as a class A amplifier for small signals. With large signals, positive excursions are limited by the saturation characteristics of the transistor (which is operated at a low collector voltage), and negative excursions generate a self-bias which drives the transistor into cutoff. Although the signal is severely distorted by this action, a tuned circuit in the output selects the FM carrier and its sidebands from the distorted signal spectrum, and thus the constant amplitude characteristic of the FM signal is restored. This is the amplitude-limiting improvement referred to previously. Only the noise phase modulation contributes to the noise at the output of the demodulator.

Amplitude limiting is also effective in reducing the interference produced by impulse-type noise, such as that generated by certain types of electrical machinery. Noise of this nature may be picked up by the antenna and superimposed as large amplitude excursions on the carrier, which the limiter removes. Limiting also can greatly alleviate the interference caused by other, weaker signals which occur within the IF bandwidth. When the limiter is either saturated or cut off by the larger signal, the weaker signal has no effect. This is known as *limiter capture* (see Young, 1990).

9.6.2 Bandwidth

When considering bandwidth, it should be kept in mind that the word is used in a number of contexts. *Signal bandwidth* is a measure of the frequency spectrum occupied by the signal. *Filter bandwidth* is the frequency range passed by circuit filters. *Channel bandwidth* refers to the overall bandwidth of the transmission channel, which in general will include a number of filters at different stages. In a well-designed system, the channel bandwidth will match the signal bandwidth.

Bandwidth requirements will be different at different points in the system. For example, at the receiver inputs for C-band and Ku-band satellite systems, the bandwidth typically is 500 MHz, accommodating

12 transponders as described in Sec. 7.7. The individual transponder bandwidth is typically 36 MHz. In contrast, the baseband bandwidth for a telephony channel is typically 3.1 kHz.

In theory, the spectrum of a frequency-modulated carrier extends to infinity. In a practical satellite system, the bandwidth of the transmitted FM signal is limited by the intermediate-frequency amplifiers. The IF bandwidth, denoted by B_{IF}, must be wide enough to pass all the significant components in the FM signal spectrum that is generated. The required bandwidth is usually estimated by *Carson's rule* as

$$B_{IF} = 2(\Delta F + F_M) \tag{9.1}$$

where ΔF is the peak carrier deviation produced by the modulating baseband signal, and F_M is the highest frequency component in the baseband signal. These maximum values, ΔF and F_M, are specified in the regulations governing the type of service. For example, for commercial FM sound broadcasting in North America, $\Delta F = 75$ kHz and $F_M = 15$ kHz.

The deviation ratio D is defined as the ratio

$$D = \frac{\Delta F}{F_M} \tag{9.2}$$

Example 9.1 A video signal of bandwidth 4.2 MHz is used to frequency modulate a carrier, the deviation ratio being 2.56. Calculate the peak deviation and the signal bandwidth.

Solution

$$D \times F = 2.56 \times 4.2 = \underline{\underline{10.752 \text{ MHz}}}$$

$$B_{IF} = 2(10.752 + 4.2) = \underline{\underline{29.9 \text{ MHz}}}$$

A similar ratio, known as the *modulation index*, is defined for sinusoidal modulation. This is usually denoted by β in the literature. Letting Δf represent the peak deviation for sinusoidal modulation and f_m the sinusoidal modulating frequency gives

$$\beta = \frac{\Delta f}{f_m} \tag{9.3}$$

The difference between β and D is that D applies for an arbitrary modulating signal and is the ratio of the maximum permitted values of deviation and baseband frequency, whereas β applies only for sinusoidal modulation (or what is often termed *tone modulation*). Very often the analysis of an FM system will be carried out for tone modulation rather than for an arbitrary signal because the mathematics is easier and the

results usually give a good indication of what to expect with an arbitrary signal.

Example 9.2 A test tone of frequency 800 Hz is used to frequency modulate a carrier, the peak deviation being 200 kHz. Calculate the modulation index and the bandwidth.

Solution

$$\beta = \frac{200}{0.8} = 250$$

$$B = 2(200 + 0.8) = \underline{\underline{401.6 \text{ kHz}}}$$

Carson's rule is widely used in practice, even though it tends to give an underestimate of the bandwidth required for deviation ratios in the range $2 < D < 10$, which is the range most often encountered in practice. For this range, a better estimate of bandwidth is given by

$$B_{\text{IF}} = 2(\Delta F + 2F_M) \tag{9.4}$$

Example 9.3 Recalculate the bandwidths for Examples 9.1 and 9.2.

Solution For the video signal,

$$B_{\text{IF}} = 2(10.75 + 8.4) = \underline{\underline{38.3 \text{ MHz}}}$$

For the 800 Hz tone:

$$B_{\text{IF}} = 2(200 + 1.6) = \underline{\underline{403.2 \text{ kHz}}}$$

In Examples 9.1 through 9.3 it will be seen that when the deviation ratio (or modulation index) is large, the bandwidth is determined mainly by the peak deviation and is given by either Eq. (9.1) or Eq. (9.4). However, for the video signal, for which the deviation ratio is relatively low, the two estimates of bandwidth are 29.9 and 38.3 MHz. In practice, the standard bandwidth of a satellite transponder required to handle this signal is 36 MHz.

The peak frequency deviation of an FM signal is proportional to the peak amplitude of the baseband signal. Increasing the peak amplitude results in increased signal power and hence a larger signal-to-noise ratio. At the same time, ΔF, and hence the FM signal bandwidth, will increase as shown previously. Although the noise power at the demodulator input is proportional to the IF filter bandwidth, the noise power output after the demodulator is determined by the bandwidth of the baseband filters, and therefore, an increase in IF filter bandwidth does not increase output noise. Thus an improvement in signal-to-noise ratio

is possible but at the expense of an increase in the IF bandwidth. This is the large-amplitude signal improvement referred to in Sec. 9.6 and considered further in the following section.

9.6.3 FM detector noise and processing gain

At the input to the FM detector, the thermal noise is spread over the IF bandwidth, as shown in Fig. 9.10*a*. The noise is represented by the system noise temperature T_s, as will be described in Sec. 12.5. At the input to the detector, the quantity of interest is the carrier-to-noise ratio. Since both the carrier and the noise are amplified equally by the receiver gain following the antenna input, this gain may be ignored in the carrier-to-noise ratio calculation, and the input to the detector represented by the voltage source shown in Fig. 9.10*b*. The carrier *root-mean-square* (rms) voltage is shown as E_c.

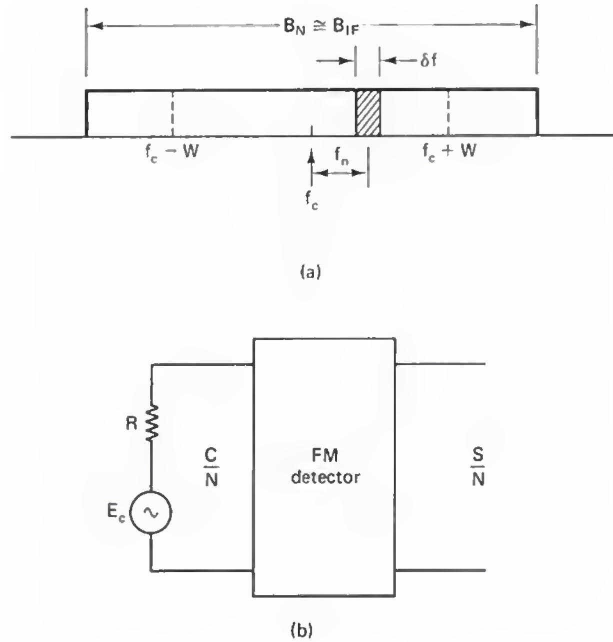


Figure 9.10 (*a*) The predetector noise bandwidth B_N is approximately equal to the IF bandwidth B_{IF}. The LF bandwidth W fixes the equivalent postdetector noise bandwidth at $2W$. δf is an infinitesimally small noise bandwidth. (*b*) Receiving system, including antenna represented as a voltage source up to the FM detector.

The available carrier power at the input to the FM detector is $E_c^2/4R$, and the available noise power at the FM detector input is kT_sB_N (as explained in Sec. 12.5), so the input carrier-to-noise ratio, denoted by C/N, is

$$\frac{C}{N} = \frac{E_c^2}{4RkT_sB_N} \tag{9.5}$$

When a sinusoidal signal of frequency, f_m, frequency modulates a carrier of frequency, f_c: The instantaneous frequency is given by $f_i = f_c + \Delta f \sin 2\pi f_m t$, where Δf is peak frequency deviation. The output signal power following the FM detector is

$$P_s = A\Delta f^2 \tag{9.6}$$

where A is a constant of the detection process.

The thermal noise at the output of a bandpass filter, for which $f_c >> B_N$ has a randomly varying amplitude component and a randomly varying phase component. (It cannot directly frequency modulate the carrier, the frequency of which is determined at the transmitter, which is at a great distance from the receiver and may be crystal controlled). When the carrier amplitude is very much greater than the noise amplitude the noise amplitude component can be ignored for FM, and the carrier angle as a function of time is $\theta(t) = 2\pi f_c t + \phi_n(t)$, where $\phi_n(t)$ is the noise phase modulation. Now the instantaneous frequency of a phase modulated wave in general is given by $\omega_i = d\theta(t)/dt$ and since $\omega_i = 2\pi f_i$, the equivalent FM resulting from the noise phase modulation is

$$f_{eq.n} = f_c + \frac{1}{2\pi}\frac{d\phi_n(t)}{dt} \tag{9.7}$$

What this shows is that the output of the FM detector, which responds to equivalent FM, is a function of the time rate of change of the phase change. Now as noted earlier, the available noise power at the input to the detector is kT_sB_N and the noise spectral density, which is the noise power per unit bandwidth just kT_s. A result from Fourier analysis is that the power spectral density of the time derivative of a waveform is $(2\pi f)^2$ times the spectral density of the input. Thus the output spectral density as a function of frequency is $(2\pi f)^2 kT_s$. The variation of output spectral noise density as a function of frequency is sketched in Fig. 9.11*a*. Since voltage is proportional to the square root of power, the noise voltage spectral density will be proportional to frequency as sketched in Fig. 9.11*b*.

Figure 9.11*a* shows that the output power spectrum is not a flat function of frequency. The available noise output power in a very small band δf would be given by $(2\pi f)^2 kT_s\delta f$. The total average noise output power

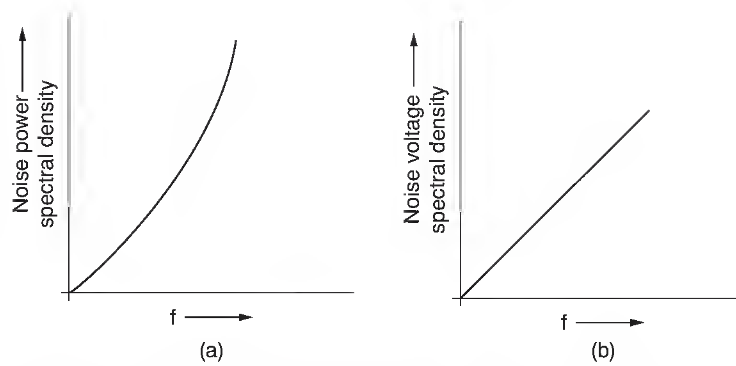


Figure 9.11 (*a*) Output noise power spectral density for FM. (*b*) The corresponding noise voltage spectral density.

would be the sum of all such increments, which is twice the area under the curve of Fig. 9.11*a*, twice because of the noise contributions from both sides of the carrier. The detailed integration required to evaluate the noise will not be carried out here, but the end result giving the signal power to noise ratio is

$$\begin{aligned}\frac{S}{N} &= \frac{P_s}{P_n} \\ &= 1.5\frac{C}{N}\frac{B_N \Delta f^2}{W^3}\end{aligned} \tag{9.8}$$

The *processing gain* of the detector is the ratio of signal-to-noise ratio to carrier-to-noise ratio. Denoting this by G_P gives

$$\begin{aligned}G_P &= \frac{S/N}{C/N} \\ &= \frac{1.5 B_N \Delta f^2}{W^3}\end{aligned} \tag{9.9}$$

Using Carson's rule for the IF bandwidth, $B_{\mathrm{IF}} = 2(\Delta f + W)$, and assuming $B_N \approx B_{IF}$, the processing gain for sinusoidal modulation becomes after some simplification

$$G_P = 3(\beta + 1)\beta^2 \tag{9.10}$$

Here, $\beta = \Delta f/W$ is the modulation index for a sinusoidal modulation frequency at the highest value W. Equation (9.10) shows that a high modulation index results in a high processing gain, which means that the signal-to-noise ratio can be increased even though the carrier-to-noise ratio is constant.

9.6.4 Signal-to-noise ratio

The term *signal-to-noise ratio* introduced in Sec. 9.6.3 is used to refer to the ratio of signal power to noise power at the receiver output. This ratio is sometimes referred to as the *postdetector* or *destination* signal-to-noise ratio. In general, it differs from the carrier-to-noise ratio at the detector input (the words *detector* and *demodulator* may be used interchangeably), the two ratios being related through the receiver processing gain as shown by Eq. (9.9). Equation (9.9) may be written in decibel form as

$$10\log_{10}\frac{S}{N} = 10\log_{10}\frac{C}{N} + 10\log_{10}G_P \tag{9.11}$$

As indicated in App. G, it is useful to use brackets to denote decibel quantities where these occur frequently. Equation (9.11) therefore may be written as

$$\left[\frac{S}{N}\right] = \left[\frac{C}{N}\right] + [G_P] \tag{9.12}$$

This shows that the signal-to-noise in decibels is proportional to the carrier-to-noise in decibels. However, these equations were developed for the condition that the noise voltage should be much less than the carrier voltage. At low carrier-to-noise ratios this assumption no longer holds, and the detector exhibits a *threshold effect.* This is a threshold level in the carrier-to-noise ratio below which the signal-to-noise ratio degrades very rapidly. The threshold level is shown in Fig. 9.12 and is defined as the carrier-to-noise ratio at which the signal-to-noise ratio is 1 dB below the straight-line plot of Eq. (9.12). For conventional FM detectors (such as the Foster Seeley detector), the threshold level may be taken as 10 dB. *Threshold extension* detector circuits are available which can provide a reduction in the threshold level of between 3 and 7 dB (Fthenakis, 1984).

In normal operation, the operating point will always be above threshold, the difference between the operating carrier-to-noise ratio and the threshold level being referred to as the *threshold margin.* This is also illustrated in Fig. 9.12.

Example 9.4 A 1-kHz test tone is used to produce a peak deviation of 5 kHz in an FM system. Given that the received [*C*/*N*] is 30 dB, calculate the receiver processing gain and the postdetector [*S*/*N*].

Solution Since the [*C*/*N*] is above threshold, Eq. (9.12) may be used. The modulation index is

$$\beta = 5\text{ kHz}/1\text{ kHz} = 5$$

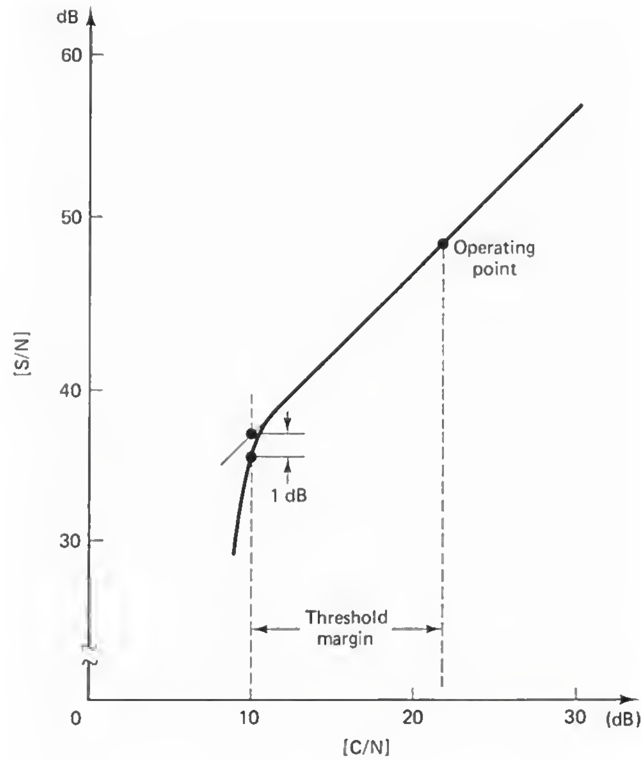


Figure 9.12 Output signal-to-noise ratio *S/N* versus input carrier-to-noise ratio *C/N* for a modulating index of 5. The straight-line section is a plot of Eq. (9.12).

Hence

$$G_P = 3 \times 5^2 \times (5 + 1) = 450$$

and

$$[G_P] = \underline{\underline{26.5 \text{ dB}}}$$

From Eq. (9.12)

$$[S/N] = 30 + 26.5 = \underline{\underline{56.5 \text{ dB}}}$$

9.6.5 Preemphasis and deemphasis

As shown in Fig. 9.11*b*, the noise voltage spectral density increases in direct proportion to the demodulated noise frequency. As a result, the signal-to-noise ratio is worse at the high-frequency end of the baseband,

a fact which is not apparent from the equation for signal-to-noise ratio, which uses average values of signal and noise power. For example, if a test tone is used to measure the signal-to-noise ratio in a TV baseband channel, the result will depend on the position of the test tone within the baseband, a better result being obtained at lower test tone frequencies. For FDM/FM telephony, the telephone channels at the low end of the FDM baseband would have better signal-to-noise ratios than those at the high end.

To equalize the performance over the baseband, a deemphasis network is introduced after the demodulator to attenuate the high-frequency components of noise. Over most of the baseband, the attenuation-frequency curve of the deemphasis network is the inverse of the rising noise-frequency characteristic shown in Fig. 9.11*b* (for practical reasons it is not feasible to have exact compensation over the complete frequency range). Thus, after deemphasis, the noise-frequency characteristic is flat, as shown in Fig. 9.13*d*. Of course, the deemphasis network also will attenuate the signal, and to correct for this, a complementary preemphasis characteristic is introduced prior to the modulator at the transmitter. The overall effect is to leave the postdetection signal levels unchanged while the high-frequency noise is attenuated. The preemphasis, deemphasis sequence is illustrated in Fig. 9.13.

The resulting improvement in the signal-to-noise ratio is referred to variously as *preemphasis improvement, deemphasis improvement,* or simply as *emphasis improvement.* It is usually denoted by P, or $[P]$ decibels, and gives the reduction in the total postdetection noise power. Preemphasis curves for FDM/FM telephony are given in CCIR Recommendation 275-2 (1978) and for TV/FM in CCIR Recommendation 405-1 (1982). CCIR values for $[P]$ are 4 dB for the top channel in multichannel telephony, 13.1 dB for 525-line TV, and 13.0 dB for 625-line TV. Taking into account the emphasis improvement, Eq. (9.12) becomes

$$\left[\frac{S}{N}\right] = \left[\frac{C}{N}\right] + [G_P] + [P] \qquad (9.13)$$

9.6.6 Noise weighting

Another factor that generally improves the postdetection signal-to-noise ratio is referred to as *noise weighting.* This is the way in which the flat-noise spectrum has to be modified to take into account the frequency response of the output device and the subjective effect of noise as perceived by the observer. For example, human hearing is less sensitive to a given noise power density at low and high audio frequencies than at the middle frequency range.

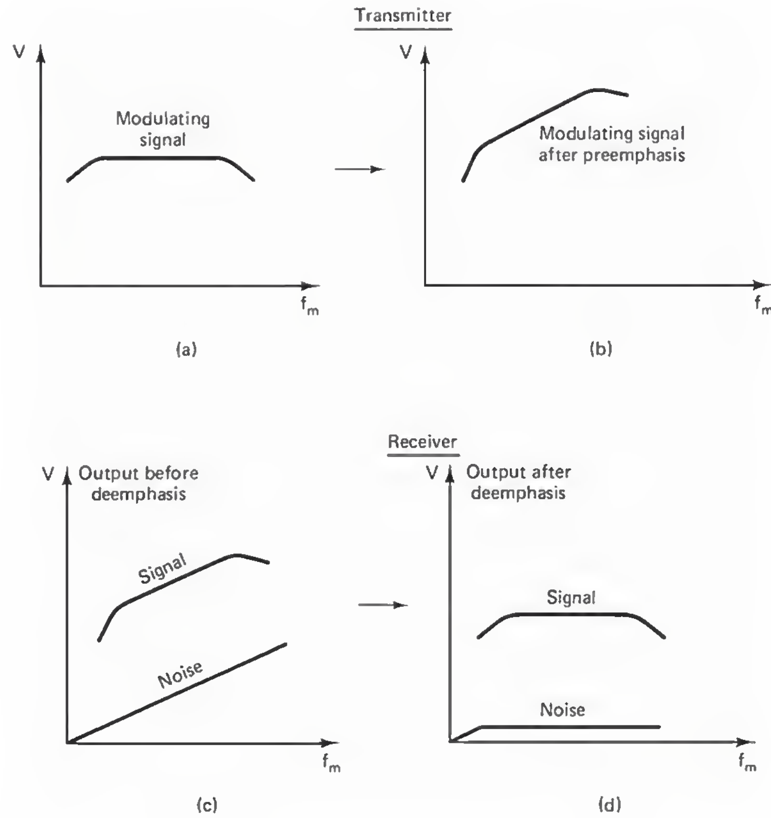


Figure 9.13 (*a* and *b*) Effect of preemphasis on the modulating signal frequency response at the transmitter. (*c* and *d*) Effect of deemphasis on the modulating signal and noise at the receiver output. The deemphasis cancels out the preemphasis for the signal while attenuating the noise at the receiver.

Weighting curves have been established for various telephone handsets in use by different telephone administrations. One of these, the CCIR curve, is referred to as the *psophometric weighting curve.* When this is applied to the flat-noise density spectrum, the noise power is reduced by 2.5 dB for a 3.1-kHz bandwidth (300–3400 Hz) compared with flat noise over the same bandwidth. The weighting improvement factor is denoted by $[W]$, and hence for the CCIR curve $[W] = 2.5$ dB. (Do not confuse the symbol W used here with that used for bandwidth earlier.) For a bandwidth of b kHz, a simple adjustment gives

$$
\begin{aligned}
[W] &= 2.5 + 10\log\frac{b}{3.1} \\
&= -2.41 + [b]
\end{aligned}
\tag{9.14}
$$

Here, b is the *numerical value of kHz* (a dimensionless number). A noise weighting factor also can be applied to TV viewing. The CCIR weighting factors are 11.7 dB for 525-line TV and 11.2 dB for 625-line TV. Taking weighting into account, Eq. (9.13) becomes

$$\left[\frac{S}{N}\right] = \left[\frac{C}{N}\right] + [G_P] + [P] + [W] \tag{9.15}$$

9.6.7 *S/N* and bandwidth for FDM/FM telephony

In the case of FDM/FM, the receiver processing gain, excluding emphasis and noise weighting, is given by (Miya, 1981, and Halliwell, 1974)

$$G_P = \frac{B_{\mathrm{IF}}}{b}\left(\frac{\Delta F_{\mathrm{rms}}}{f_m}\right)^2 \tag{9.16}$$

Here, f_m is a specified baseband frequency in the channel of interest, at which G_P is to be evaluated. For example, f_m may be the center frequency of a given channel, or it may be the top frequency of the baseband signal. The channel bandwidth is b (usually 3.1 kHz), and ΔF_{rms} is the root-mean-square deviation per channel of the signal. The rms deviation is determined under specified test tone conditions, details of which will be found in CCIR Recommendation 404-2 (1982). Some values are shown in Table 9.1.

Because ΔF_{rms} is determined for a test tone modulation, the peak deviation for the FDM waveform has to take into account the waveform shape through a factor g. This is a voltage ratio that is usually expressed in decibels. For a small number of channels, g may be as high as 18.6 dB (Ffthenakis, 1984), and typical values range from 10 to 13 dB. For the number of channels n greater than 24, the value of 10 dB is often

TABLE 9.1 FDM/FM RMS Deviations

Maximum number of channels	RMS deviations per channel (kHz)
12	35
24	35
60	50, 100, 200
120	50, 100, 200
300	200
600	200
960	200
1260	140, 200
1800	140
2700	140

used. Denoting the decibel value as gdB, then the voltage ratio is obtained from

$$g = 10^{\text{gdB}/20} \tag{9.17}$$

The peak deviation also will depend on the number of channels, and this is taken into account through use of a *loading factor*, L. The relevant CCITT formulas are

$$\text{For } n > 240\text{: } 20\log L = -15 + 10\log n \tag{9.18}$$

$$\text{For } 12 \le n \le 240\text{: } 20\log L = -1 + 4\log n \tag{9.19}$$

Once L and g are found, the required peak deviation is obtained from the tabulated rms deviation as

$$\Delta F = g \cdot L \cdot \Delta F_{\text{rms}} \tag{9.20}$$

The required IF bandwidth can now be found using Carson's rule, Eq. (9.1), and the processing gain from Eq. (9.16). The following example illustrates the procedure.

Example 9.5 The carrier-to-noise ratio at the input to the demodulator of an FDM/FM receiver is 25 dB. Calculate the signal-to-noise ratio for the top channel in a 24-channel FDM baseband signal, evaluated under test conditions for which Table 9.1 applies. The emphasis improvement is 4 dB, noise weighting improvement is 2.5 dB, and the peak/rms factor is 13.57 dB. The audio channel bandwidth may be taken as 3.1 kHz.

Solution Given data: $n = 24$; $g\text{dB} = 13.57$; $b = 3.1$ kHz; $[P] = 4$; $[W] = 2.5$; $[C/N] = 25$

From Eq. (9.17):

$$g = 10^{\text{gdB}/20} = 4.77$$

From Eq. (9.19):

$$L = 10^{(-1+4\log n)/20} = 1.683$$

From Table 9.1, for 24 channels $\Delta F_{\text{rms}} = 35$ kHz, and using Eq. (9.20),

$$\Delta F = g \cdot L \cdot \Delta F_{\text{rms}} \cong 281 \text{ kHz}$$

Assuming that the baseband spectrum is as shown in Fig. 9.4*a*, the top frequency is

$$f_m = 108 \text{ kHz}$$

and Carson's rule gives

$$B_{\text{IF}} = 2(\Delta F + f_m) = 778 \text{ kHz}$$

From Eq. (9.16):

$$G_P = \frac{777.8}{3.1}\left(\frac{35}{108}\right)^2 \cong 26.36$$

From Eq. (9.15):

$$\left[\frac{S}{N}\right] = \left[\frac{C}{N}\right] + 10\log G_P + [P] + [W]$$
$$= 25 + 14.21 + 4 + 2.5$$
$$= \underline{\underline{45.7\text{ dB}}}$$

9.6.8 Signal-to-noise ratio for TV/FM

Television performance is measured in terms of the postdetector video signal-to-noise ratio, defined as (CCITT Recommendation 567-2, 1986)

$$\left(\frac{S}{N}\right)_V = \frac{\text{peak-to-peak video voltage}}{\text{rms noise voltage}} \tag{9.21}$$

Because peak-to-peak video voltage is used, $2\Delta F$ replaces ΔF in Eq. (9.8). Also, since power is proportional to voltage squared,

$$\left(\frac{S}{N}\right)_v^2 = 1.5\frac{C}{N}\frac{B_N(2\Delta F)^2}{W^3} \tag{9.22}$$

where W is the highest video frequency. With the deviation ratio $D = \Delta F/W$, and the processing gain for TV denoted as G_{PV},

$$\begin{aligned} G_{\text{PV}} &= \frac{(S/N)_V^2}{C/N} \\ &= 12D^2(D + 1) \end{aligned} \tag{9.23}$$

Some workers include an implementation margin to allow for nonideal performance of filters and demodulators (Bischof et al., 1981).

With the implementation margin in decibels denoted by [IMP], Eq. (9.15) becomes

$$\left[\left(\frac{S}{N}\right)_V^2\right] = \left[\frac{C}{N}\right] + [G_{\text{PV}}] + [P] + [W] - [\text{IMP}] \tag{9.24}$$

Recall that the square brackets denote decibels, that is, $[X] = 10\log_{10}X$. This is illustrated in the following example.

Example 9.6 A satellite TV link is designed to provide a video signal-to-noise ratio of 62 dB. The peak deviation is 9 MHz, and the highest video baseband frequency is 4.2 MHz. Calculate the carrier-to-noise ratio required at the input to the FM detector, given that the combined noise weighting, emphasis improvement, and implementation margin is 11.8 dB.

Solution

$$D = \frac{9}{4.2} = 2.143$$

Equation (9.23) gives:

$$G_{PV} = 12 \times 2.143^2 \times (2.143 + 1) = 173.2$$

Therefore,

$$[G_{PV}] = 10 \log 173.2 = 22.4 \text{ dB}$$

Since the required signal-to-noise ratio is 62 dB, Eq. (9.24) can be written as

$$62 = \left[\frac{C}{N}\right] + 22.4 + 11.8$$

from which $[C/N] = \underline{\underline{27.8 \text{ dB}}}$.

9.7 Problems and Exercises

9.1. State the frequency limits generally accepted for telephone transmission of speech and typical signal levels encountered in the telephone network.

9.2. Show that when two sinusoids of different frequencies are multiplied together, the resultant product contains sinusoids at the sum and difference frequencies only. Hence show how a multiplier circuit may be used to produce a DSBSC signal.

9.3. Explain how a DSBSC signal differs from a conventional amplitude modulated signal such as used in the medium-wave (broadcast) radio band. Describe one method by which an SSB signal may be obtained from a DSBSC signal.

9.4. Explain what is meant by FDM telephony. Sketch the frequency plans for the CCITT designations of group, supergroup, basic mastergroup, and super mastergroup.

9.5. With the aid of a block schematic, show how 12 VF channels could be frequency-division multiplexed.

9.6. Explain how a 252-VF-channel group is formed for satellite transmission.

9.7. Describe the essential features of the video signal used in the NTSC color TV scheme. How is the system made compatible with monochrome reception?

9.8. Explain how the sound information is added to the video information in a color TV transmission.

9.9. For the matrix network *M* shown in Fig. 9.7, derive the equations for the Y, Q, and I signals in terms of the input signals R, G, and B.

9.10. Explain what is meant by *frequency modulation*. A 70-MHz carrier is frequency modulated by a 1-kHz tone of 5-V peak amplitude. The frequency deviation constant is 15 kHz/V. Write down the expression for instantaneous frequency.

9.11. An angle-modulated wave may be written as $\sin\theta(t)$, where the argument $\theta(t)$ is a function of the modulating signal. Given that the instantaneous angular frequency is $\omega_i = d\theta(t)/dt$, derive the expression for the FM carrier in Prob. 9.10.

9.12. (*a*) Explain what is meant by *phase modulation*. (*b*) A 70-MHz carrier is phase modulated by a 1-kHz tone of 5-V peak amplitude. The phase modulation constant is 0.1 rad/V. Write down the expression for the argument $\theta(t)$ of the modulated wave.

9.13. Determine the equivalent peak frequency deviation for the phase-modulated signal of Prob. 9.12.

9.14. Show that when a carrier is phase modulated with a sinusoid, the equivalent peak frequency deviation is proportional to the modulating frequency. Explain the significance of this on the output of an FM receiver used to receive the PM wave.

9.15. In the early days of FM it was thought that the bandwidth could be limited to twice the peak deviation irrespective of the modulating frequency. Explain the fallacy behind this reasoning.

9.16. A 10-kHz tone is used to frequency modulate a carrier, the peak deviation being 75 kHz. Use Carson's rule to estimate the bandwidth required.

9.17. A 70-MHz carrier is frequency modulated by a 1-kHz tone of 5-V peak amplitude. The frequency deviation constant is 15 kHz/V. Use Carson's rule to estimate the bandwidth required.

9.18. A 70-MHz carrier is phase modulated by a 1-kHz tone of 5-V peak amplitude. The phase modulation constant is 0.1 rad/V. Find the equivalent peak deviation and, hence, use Carson's rule to estimate the bandwidth required for the PM signal.

9.19. Explain what is meant by *preemphasis* and *deemphasis* and why these are effective in improving signal-to-noise ratio in FM transmission. State typical improvement levels expected for both telephony and TV transmissions.

9.20. Explain what is meant by *noise weighting*. State typical improvement levels in signal-to-noise ratios which result from the introduction of noise weighting for both telephony and TV transmissions.

9.21. Calculate the loading factor L for (*a*) a 12-channel, (*b*) a 120-channel, and (*c*) an 1800-channel FDM/FM telephony signal.

9.22. Calculate the IF bandwidth required for (*a*) a 12-channel, (*b*) a 300-channel, and (*c*) a 960-channel FDM/FM telephony signal. Assume that the peak/rms factor is equal to 10 dB for parts (*a*) and (*b*) and equal to 18 dB for part (*c*).

9.23. Calculate the receiver processing gain for each of the signals given in Prob. 9.22.

9.24. A video signal has a peak deviation of 9 MHz and a video bandwidth of 4.2 MHz. Using Carson's rule, calculate the IF bandwidth required and the receiver processing gain.

9.25. For the video signal of Prob. 9.24, the emphasis improvement figure is 13 dB, and the noise weighting improvement figure is 11.2 dB. Calculate in decibels (*a*) the signal-to-noise power ratio and (*b*) the video signal-to-noise ratio as given by Eq. (9.21). The $[C/N]$ value is 22 dB. Assume a sinusoidal video signal.

References

Bischof, I. J., W. B. Day, R. W. Huck, W. T. Kerr, and N. G. Davies. 1981. "Anik-B Program Delivery Pilot Project. A 12-month Performance Assessment." CRC Report No. 1349, Dept. of Communications, Ottawa, December.

Campanella, S. J. 1983. *Companded Single Sideband (CSSB) AM/FDMA Performance*. Wiley, New York.

CCIR Recommendation 275-2. 1978. "Pre-emphasis Characteristic in Frequency Modulation Radio Relay Systems for Telephony Using Frequency-Division Multiplex." *14th Plenary Assembly*, Vol. IX, Kyoto.

CCIR Recommendation 404-2. 1982. "Frequency Division for Analog Radio Relay Systems for Telephony Using Frequency Division Multiplex." *15th Plenary Assembly*, Vol. IX, part 1, Geneva.

CCIR Recommendation 405-1. 1982. "Pre-emphasis Characteristics for Frequency Modulation Radio Relay Systems for Television." *15th Plenary Assembly*, Vol. IX, Part 1, Geneva.

CCITT G423. 1976. "Interconnection at the Baseband Frequencies of Frequency-Division Multiplex Radio-Relay Systems 1, 2." *International Carrier Analog Systems,* Vol. III, Part 2, Geneva.

CCITT Recommendation 567-2. 1986. "Transmission Performance of Television Circuits Designed for Use in International Circuits." Vol. XII, Geneva.

CCITT Recommendation G322. 1976. "General Characteristics Recommended for Systems on Symmetric Pair Cables." *International Carrier Analog Systems*, Vol. III, Part 2, Geneva.

Freeman, Roger L. 1981. *Telecommunications Systems Engineering.* Wiley, New York.

Fthenakis, Emanuel. 1984. *Manual of Satellite Communications.* McGraw-Hill, New York.

Halliwell, B. J. (ed.). 1974. *Advanced Communication Systems.* Newnes-Butterworths, London.

Miya, K. (ed.). 1981. *Satellite Communications Technology.* KDD Engineering and Consulting, Tokyo, Japan.

Young, P. H. 1990. *Electronic Communication Techniques.* Merrill Publishing Company, New York.

Chapter

10

Digital Signals

10.1 Introduction

As already mentioned in connection with analog signals, baseband signals are those signals which occupy the lowest, or base, frequency band in the frequency spectrum used by the telecommunications network. A baseband signal may consist of one or more information signals.

For example, a number of telephony signals in digital form may be combined into one baseband signal by the process known as *time-division multiplexing*.

Analog signals may be converted into digital signals for transmission. Digital signals also originate in the form of computer and other data. In general, a digital signal is a coded version of the original data or analog signal. In this chapter, the characteristics of the more common types of digital baseband signals are described, along with representative methods of digital modulation.

10.2 Digital Baseband Signals

Digital signals are coded representations of information. Keyboard characters, for example, are usually encoded in binary digital code. A *binary code* has two symbols, usually denoted as 0 and 1, and these are combined to form binary words to represent the characters. For example, a teleprinter code may use the combination 11000 to represent the letter A.

Analog signals such as speech and video may be converted to a digital form through an *analog-to-digital* (A/D) converter. A particular form of A/D conversion is employed, known as *pulse-code modulation*, which will be described in detail later. Some of these sources are illustrated diagrammatically in Fig. 10.1.

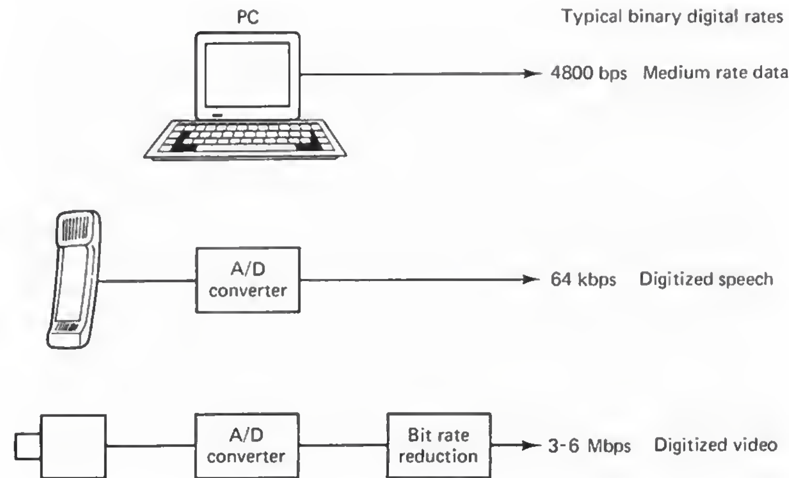


Figure 10.1 Examples of binary data sources.

In *digital* terminology, a binary symbol is known as a *binit* from *bin*ary dig*it*. The *information* carried by a binit is, in most practical situations, equal to a unit of information known as a *bit*. Thus it has become common practice to refer to binary symbols as bits rather than binits, and this practice will be followed here.

The *digital* information is transmitted as a waveform, some of the more common waveforms used for binary encoding being shown in Fig. 10.2. These will be referred to as *digital waveforms,* although strictly speaking they are analog representations of the digital information being transmitted. The binary sequence shown in Fig. 10.2 is 1010111. Detailed reasons for the use of different waveforms will be found in most books on digital communications (see Bellamy, 1982).

The duration of a bit is referred to as the *bit period* and is shown as T_b. The bit rate is given by

$$R_b = \frac{1}{T_b} \tag{10.1}$$

With T_b in seconds, the bit rate will be in bits per second, usually denoted by b/s.

Figure 10.2*a* shows a *unipolar* waveform, meaning that the waveform excursions from zero are always in the same direction, either positive or negative. They are shown as positive A in Fig. 10.2*a*. Because it has a dc component, the unipolar waveform is unsuitable for use on telephone lines and radio networks, including satellite links.

Figure 10.2*b* shows a *polar* waveform, which utilizes positive and negative polarities. (In Europe this is referred to as a *bipolar* waveform, but

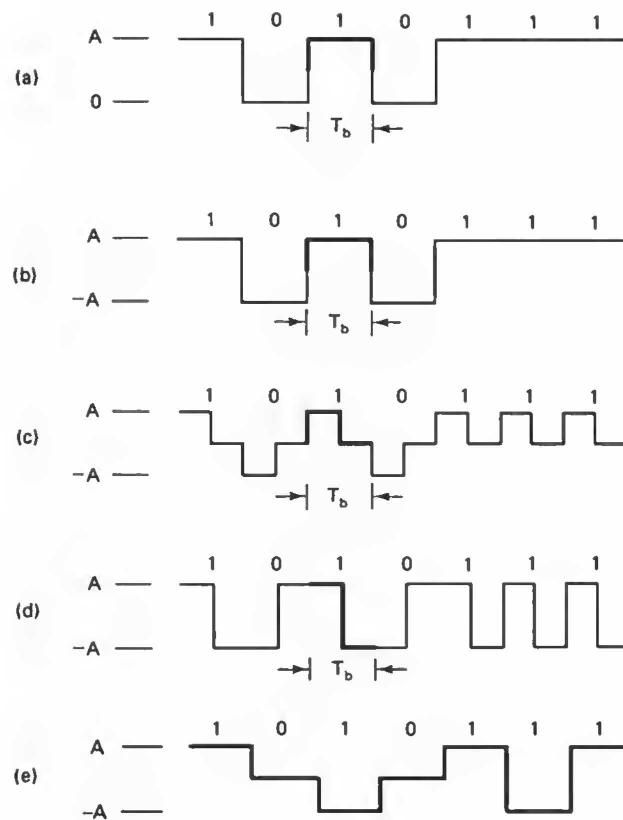


Figure 10.2 Examples of binary waveforms used for encoding digital data: (*a*) unipolar NRZ; (*b*) polar NRZ; (*c*) polar RZ; (*d*) split phase or Manchester; (*e*) alternate mark inversion (AMI).

the term *bipolar* in North American usage is reserved for a specific waveform, described later). For a long, random sequence of 1s and 0s, the dc component would average out to zero. However, long sequences of like symbols result in a gradual drift in the dc level, which creates problems at the receiver decoder. Also, the decoding process requires knowledge of the bit timing, which is derived from the zero crossovers in the waveform, and these are obviously absent in long strings of like symbols. Both the unipolar and polar waveforms shown in Fig. 10.2*a* and *b* are known as *non-return-to-zero* (NRZ) *waveforms*. This is so because the waveform does not return to the zero baseline at any point during the bit period.

Figure 10.2*c* shows an example of a polar *return-to-zero* (RZ) *waveform*. Here, the waveform does return to the zero baseline in the middle of the bit period, so transitions will always occur even within a long string of

like symbols, and bit timing can be extracted. However, dc drift still occurs with long strings of like symbols.

In the *split-phase* or *Manchester encoding* shown in Fig. 10.2*d*, a transition between positive and negative levels occurs in the middle of each bit. This ensures that transitions will always be present so that bit timing can be extracted, and because each bit is divided equally between positive and negative levels, there is no dc component.

A comparison of the frequency bandwidths required for digital waveforms can be obtained by considering the waveforms which alternate at the highest rate between the two extreme levels. These will appear as squarewaves. For the basic polar NRZ waveform of Fig. 10.2*b*, this happens when the sequence is . . . 101010 . . . The periodic time of such a squarewave is $2T_b$, and the fundamental frequency component is $1/2T_b$. For the split-phase encoding, the squarewave with the highest repetition frequency occurs with a long sequence of like symbols such as . . . 1111111 . . ., as shown in Fig. 10.2*d*. The periodic time of this squarewave is T_b, and hence the fundamental frequency component is twice that of the basic polar NRZ. Thus the split-phase encoding requires twice the bandwidth compared with that for the basic polar NRZ, while the bit rate remains unchanged. The utilization of bandwidth, measured in bits per second per hertz, is therefore less efficient.

An *alternate mark inversion* (AMI) *code* is shown in Fig. 10.2*e*. Here, the binary 0s are at the zero baseline level, and the binary 1s alternate in polarity. In this way, the dc level is removed, while bit timing can be extracted easily, except when a long string of zeros occurs. Special techniques are available to counter this last problem. The highest pulse-repetition frequency occurs with a long string of . . . 111111 . . . the periodic time of which is $2T_b$, the same as the waveform of Fig. 10.2*b*. The AMI waveform is also referred to as a *bipolar* waveform in North America.

Bandwidth requirements may be reduced by utilizing multilevel digital waveforms. Figure 10.3*a* shows a polar NRZ signal for the sequence 11010010. By arranging the bits in groups of two, four levels can be used. For example, these may be

11	3A
10	A
01	–A
00	–3A

This is referred to as *quaternary* encoding, and the waveform is shown in Fig. 10.3*b*. The encoding is symmetrical about the zero axis, the spacing between adjacent levels being $2A$. Each level represents a *symbol*, the duration of which is the *symbol period*. For the quaternary waveform the

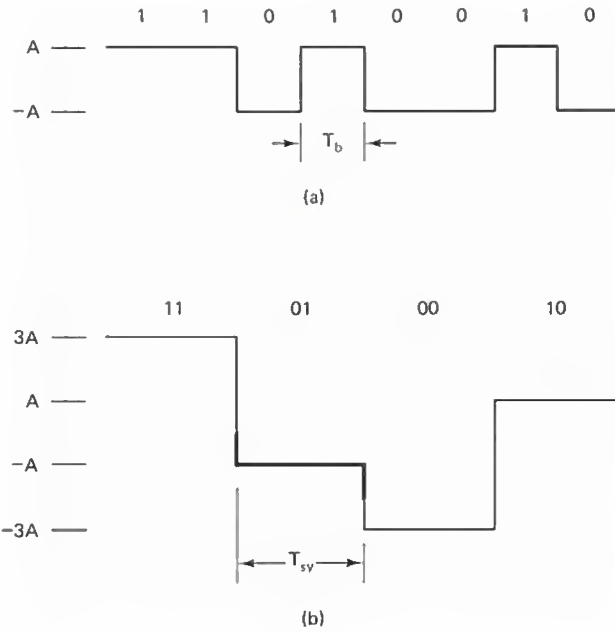


Figure 10.3 Encoding of 11010010 in (*a*) binary polar NRZ and (*b*) quaternary polar NRZ.

symbol period is seen to be equal to twice the bit period, and the symbol rate is

$$R_{\text{sym}} = \frac{1}{T_{\text{sym}}} \tag{10.2}$$

The symbol rate is measured in units of *bauds*, where 1 Bd is one symbol per second.

The periodic time of the squarewave having the greatest symbol repetition frequency is $2T_{\text{sym}}$, which is equal to $4T_b$, and hence the bandwidth, compared with the basic binary waveform, is halved. The bit rate (as distinct from the symbol rate) remains unchanged, and hence the bandwidth utilization in terms of bits per second per hertz is doubled.

In general, a waveform may have M levels (sometimes referred to as an *M-ary waveform*), where each symbol represents m bits and

$$m = \log_2 M \tag{10.3}$$

The symbol period is therefore

$$T_{\text{sym}} = mT_b \tag{10.4}$$

and the symbol rate in terms of bit rate is

$$R_{\text{sym}} = \frac{R_b}{m} \tag{10.5}$$

For satellite transmission, the encoded message must be modulated onto the microwave carrier. Before examining the modulation process, we describe the way in which speech signals are converted to a digital format through pulse code modulation.

10.3 Pulse Code Modulation

In the previous section describing baseband digital signals, the information was assumed to be encoded in one of the digital waveforms shown in Figs. 10.2 and 10.3. Speech and video appear naturally as analog signals, and these must be converted to digital form for transmission over a digital link. In Fig. 10.1 the speech and video analog signals are shown converted to digital form through the use of A/D converters. The particular form of A/D conversion used is known as *pulse-code modulation* (PCM). Commercially available integrated circuits known as PCM *codecs* (for coder-decoder) are used to implement PCM. Figure 10.4*a* shows a block schematic for the Motorola MC145500 series of codecs suitable for speech signals. The analog signal enters at the Tx terminals and passes through a low-pass filter, followed by a high-pass filter to remove any 50/60-Hz interference which may appear on the line. The low-pass filter has a cutoff frequency of about 4 kHz, which allows for the filter rolloff above the audio limit of 3400 Hz. As shown in connection with single-sideband systems, a voice channel bandwidth extending from 300 to 3400 Hz is considered satisfactory for speech. Band limiting the audio signal in this way reduces noise. It has another important consequence associated with the analog-to-digital conversion process. The analog signal is digitized by taking samples at periodic intervals. A theorem, known as the *sampling theorem*, states in part that the *sampling frequency* must be at least twice the highest frequency in the spectrum of the signal being sampled. With the upper cutoff frequency of the audio filter at 4 kHz, the sampling frequency can be standardized at 8 kHz.

The sampled voltage levels are encoded as binary digital numbers in the A/D converter following the high-pass filter. The binary number which is transmitted actually represents a range of voltages, and all samples which fall within this range are encoded as the same number. This process, referred to as *quantization*, obviously will introduce some distortion (termed *quantization noise*) into the signal. In a properly designed system, the quantization noise is kept well within acceptable limits. The quantization steps follow a nonlinear law, with



(a)

Chord Number	Number of Steps	Step Size	Normalized Encode Decision Levels	Digital Code 1 Sign	2 Chord	3 Chord	4 Chord	5 Step	6 Step	7 Step	8 Step	Normalized Decode Levels
6	16	256	8159	1	0	0	0	0	0	0	0	8031
			7903 : 4319				:					:
				1	0	0	0	1	1	1	1	4191
7	16	128	4063 : 2143				:					:
				1	0	0	1	1	1	1	1	2079
6	16	64	2015 : 1055				:					:
				1	0	1	0	1	1	1	1	1023
5	16	32	991 : 511				:					:
				1	0	1	1	1	1	1	1	495
4	16	16	479 : 239				:					:
				1	1	0	0	1	1	1	1	231
3	16	8	223 : 103				:					:
				1	1	0	1	1	1	1	1	99
2	16	4	95 : 35				:					:
				1	1	1	0	1	1	1	1	33
1	15	2	31 : 3				:					:
				1	1	1	1	1	1	1	0	2
	1	1	1	1	1	1	1	1	1	1	1	0
			0									

NOTES:
1. Characteristics are symmetrical about analog zero with sign bit = 0 for negative analog values.
2. Digital code includes inversion of all magnitude bits.

(b)

Figure 10.4 (*a*) MC145500/01/02/03/05 PCM CODEC/filter monocircuit block diagram. (*b*) μ-law encode-decode characteristics. (*Courtesy of Motorola, Inc.*)

large signals being quantized into coarser steps than small signals. This is termed *compression*, and it is introduced to keep the signal-to-quantization noise ratio reasonably constant over the full dynamic range of the input signal while maintaining the same number of bits per codeword. At the receiver (the D/A block in Fig. 10.4), the binary codewords are automatically decoded into the larger quantized steps for the larger signals, this being termed *expansion*. The expansion law is the inverse of the compression law, and the combined processing is termed *companding*.

Figure 10.4*b* shows how the MC145500 codec achieves compression by using a *chorded approximation*. The leading bit of the digital codeword is a *sign* bit, being 1 for positive and 0 for negative samples of the analog signal. The next three bits are used to encode the chord in which the analog signal falls, the three bits giving a total of eight chords. Each chord is made to cover the same number of input steps, but the step size increases from chord to chord. The chord bits are followed by four bits indicating the step in which the analog value lies. The normalized decision levels shown in Fig. 10.4*b* are the analog levels at which the comparator circuits change from one chord to the next and from one step to the next. These are normalized to a value 8159 for convenience in presentation. For example, the maximum value may be considered to be 8159 mV, and then the smallest step would be 1 mV. The first step is shown as 1 (mV), but it should be kept in mind that the first quantized level spans the analog zero so that 0^+ must be distinguished from 0^-. Thus the level representing zero has in fact a step size of ± 1 mV.

As an example, suppose the sampled analog signal has a value +500 mV. This falls within the normalized range 479 to 511 mV, and therefore, the binary code is 10111111. It should be mentioned that normally the first step in a chord would be encoded 0000, but the bits are inverted, as noted in Fig. 10.4*b*. This is so because low values are more likely than high values, and inversion increases the 1-bit density, which helps in maintaining synchronization.

The table in Fig. 10.4*b* shows mu-law encode-decode characteristics. The term *mu law*, usually written as μ-law, originated with older analog compressors, where μ was a parameter in the equation describing the compression characteristic. The μ-law characteristic is standard in North America and Japan, while in Europe and many other parts of the world a similar law known as the *A-law* is in use. Figure 10.5 shows the curves for $\mu = 255$ and $A = 87.6$, which are the standard values in use. These are shown as smooth curves, which could be approached with the older analog compression circuits. The chorded approximation approaches these in straight-line segments, or *chords*, for each step.

Because of the similarity of the A-law and μ-law curves, the speech quality, as affected by companding, will be similar in both systems, but

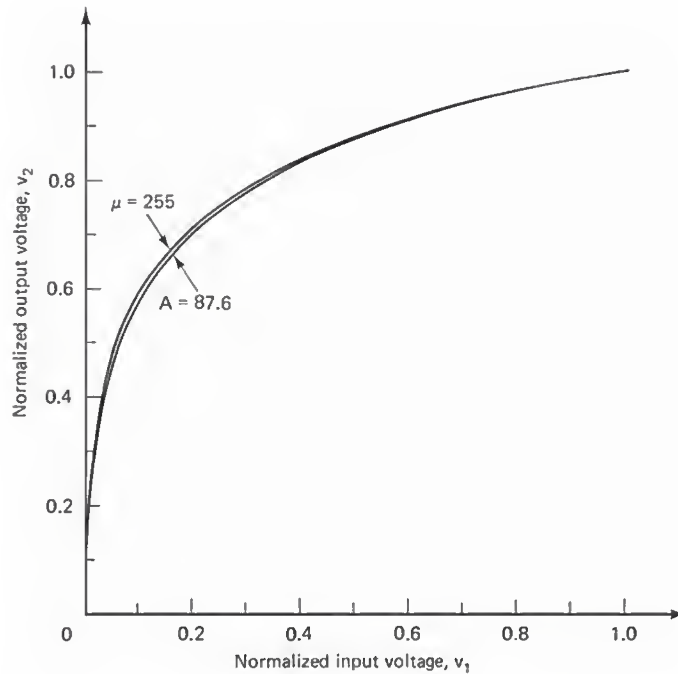


Figure 10.5 Compressor characteristics. Input and output voltage scales are normalized to the maximum values.

otherwise the systems are incompatible, and conversion circuitry is required for interconnections such as might occur with international traffic. The MC145500 can be configured for use with either law through appropriate pin selections, but of course the transmitting and receiving functions must be configured for the same law.

In the receiver, the output from the D/A converter is passed through a low-pass filter which selects the original analog spectrum from the quantized signal. Its characteristics are similar to those of the low-pass filter used in the transmitter. Apart from the quantization noise (which should be negligible), the final output is a replica of the filtered analog signal at the transmitter.

With a sampling rate of 8 kHz or 8000 samples per second and 8 bits for each sample codeword, the bit rate for a single-channel PCM signal is

$$R_b = 8000 \times 8 = 64 \text{ kb/s} \tag{10.6}$$

The frequency spectrum occupied by a digital signal is proportional to the bit rate, and in order to conserve bandwidth, it may be necessary to reduce the bit rate. For example, if 7-bit codewords were to be used,

the bit rate would be 56 kb/s. Various data reduction schemes are in use which give much greater reductions, and some of these can achieve bit rates as low as 2400 b/s (Hassanein et al., 1989 and 1992).

10.4 Time-Division Multiplexing

A number of signals in binary digital form can be transmitted through a common channel by interleaving the pulses in time, this being referred to as *time-division multiplexing* (TDM). For speech signals, a separate codec may be used for each voice channel, the outputs from these being combined to form a TDM baseband signal, as shown in Fig. 10.6. At the baseband level in the receiver, the TDM signal is demultiplexed, the PCM signals being routed to separate codecs for decoding. In satellite systems, the TDM waveform is used to modulate the carrier wave, as described later.

The time-division multiplexed signal format is best described with reference to the widely used Bell T1 system. The signal format is illustrated in Fig. 10.7*a*. Each PCM word contains 8 bits, and a *frame* contains 24 PCM channels. In addition, a periodic *frame synchronizing* signal must be transmitted, and this is achieved by inserting a bit from the frame synchronizing codeword at the beginning of every frame. At the receiver, a special detector termed a *correlator* is used to detect the frame synchronizing codeword in the bit stream, which enables the frame timing to be established. The total number of bits in a frame is therefore $24 \times 8 + 1 = 193$. Now, as established earlier, the sampling frequency for voice is 8 kHz, and so the interval between PCM words for a given channel is $1/8000 = 125\ \mu s$. For example, the leading bit in the PCM codewords

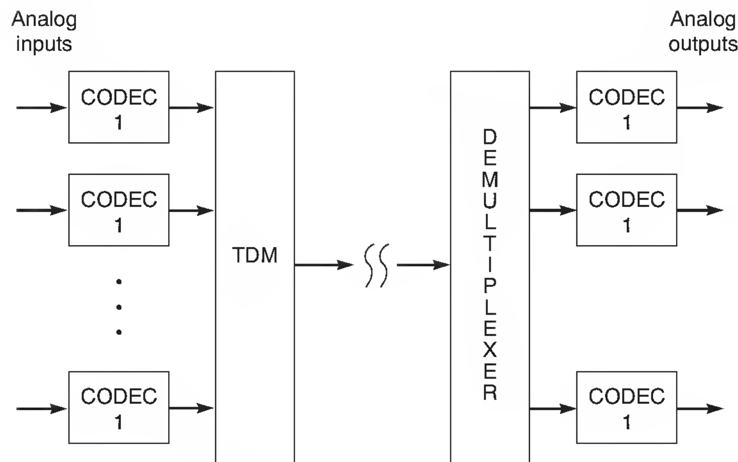


Figure 10.6 A basic TDM system.

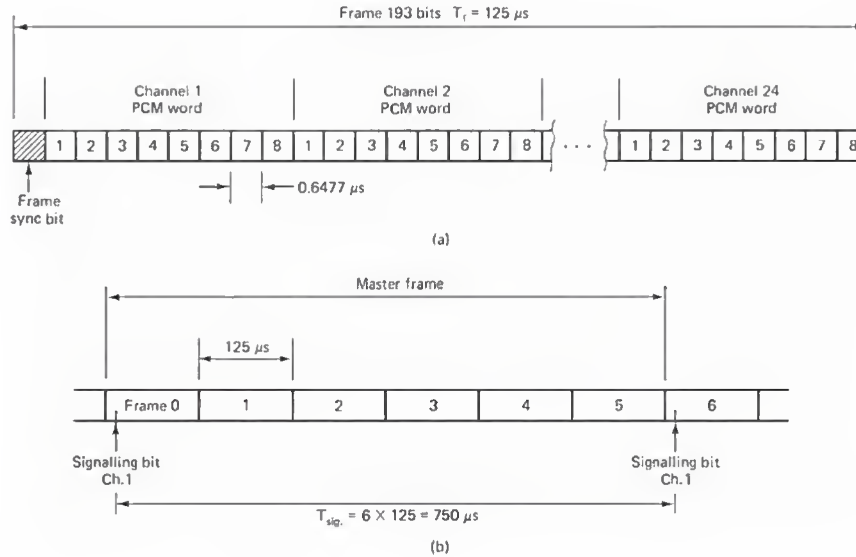


Figure 10.7 Bell T1 PCM format.

for a given channel must be separated in time by no more than 125 μs. As can be seen from Fig. 10.7*a*, this is also the frame period, and therefore, the bit rate for the T1 system is

$$R_b = \frac{193}{125 \times 10^{-6}} = 1.544 \text{ Mb/s} \tag{10.7}$$

Signaling information is also carried as part of the digital stream. *Signaling* refers to such data as number dialed, busy signals, and billing information. Signaling can take place at a lower bit rate, and in the T1 system, the eighth bit for every channel, in every sixth frame, is replaced by a signaling bit. This is illustrated in Fig. 10.7*b*. The time separation between signaling bits is 6 × 125 μs = 750 μs, and the signaling bit rate is therefore 1/(750 μs) = 1.333 kb/s.

10.5 Bandwidth Requirements

In a satellite transmission system, the baseband signal is modulated onto a carrier for transmission. Filtering of the signals takes place at a number of stages. The baseband signal itself is band-limited by filtering to prevent the generation of excessive sidebands in the modulation process. The modulated signal undergoes *bandpass filtering* (BPF) as part of the amplification process in the transmitter.

Where transmission lines form the channel, the frequency response of the lines also must be taken into account. With a satellite link, the main channel is the radiofrequency path, which has little effect on the frequency spectrum but does introduce a propagation delay which must be taken into account.

At the receive end, bandpass filtering of the incoming signal is necessary to limit the noise which is introduced at this stage. Thus the signal passes through a number of filtering stages, and the effect of these on the digital waveform must be taken into account.

The spectrum of the output pulse at the receiver is determined by the spectrum of the input pulse $V_i(f)$, the transmit filter response $H_T(f)$, the channel frequency response $H_{\mathrm{CH}}(f)$, and the receiver filter response $H_R(f)$. These are shown in Fig. 10.8. Thus

$$V(f) = V_i(f)H_T(f)H_{\mathrm{CH}}(f)H_R(f) \tag{10.8}$$

Inductive and capacitive elements are an inherent part of the filtering process. These do not dissipate power, but energy is periodically cycled between the magnetic and electric fields and the signal. The time required for this energy exchange results in part of the signal being delayed so that a square pulse entering at the transmitting end may exhibit "ringing" as it exits at the receiving end. This is illustrated in Fig. 10.9*a*.

Because the information is digitally encoded in the waveform, the distortion apparent in the pulse shape is not important as long as the receiver can distinguish the binary 1 pulse from the binary 0 pulse. This requires the waveform to be sampled at the correct instants in order to determine its polarity. With a continuous waveform, the "tails" which result from the "ringing" of all the preceding pulses can combine to interfere with the particular pulse being sampled. This is known as *intersymbol interference* (ISI), and it can be severe enough to produce an error in the detected signal polarity.

The ringing cannot be removed, but the pulses can be shaped such that the sampling of a given pulse occurs when the tails are at zero crossover points. This is illustrated in Fig. 10.9*b*, where two tails are shown overlapping the pulse being sampled. In practice, perfect pulse shaping cannot be achieved, so some ISI occurs, but it can be reduced to negligible proportions.

The pulse shaping is carried out by controlling the spectrum of the received pulse as given by Eq. (10.8). One theoretical model for the

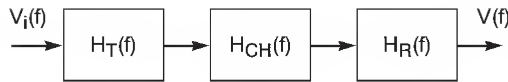
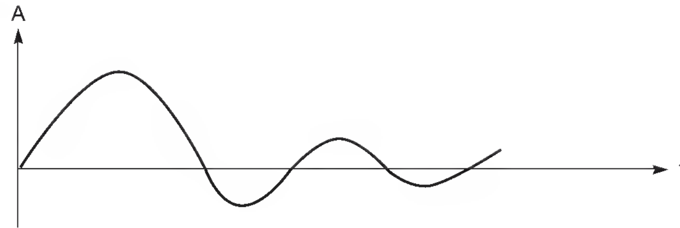


Figure 10.8 Frequency spectrum components of Eq. (10.8).



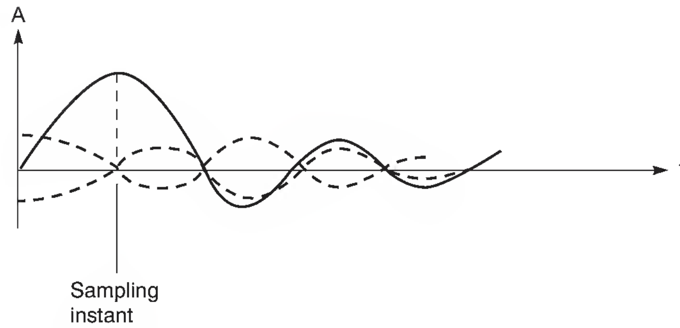


Figure 10.9 (*a*) Pulse ringing. (*b*) Sampling to avoid ISI.

spectrum is known as the *raised cosine response*, which is shown in Fig. 10.10. Although a theoretical model, it can be approached closely with practical designs. The raised cosine spectrum is described by

$$V(f) = \begin{cases} 1 & \text{for } f < f_1 \\ 0.5\left(1 + \cos\dfrac{\pi(f - f_1)}{B - f_1}\right) & \text{for } f_1 < f < B \\ 0 & \text{for } B < f \end{cases} \tag{10.9}$$

The frequencies f_1 and B are determined by the symbol rate and a design parameter known as the *rolloff factor*, denoted here by the symbol ρ. The rolloff factor is a specified parameter in the range

$$0 \leq \rho \leq 1 \tag{10.10}$$

In terms of ρ and the symbol rate, the bandwidth B is given by

$$B = \frac{1 + \rho}{2} R_{\text{sym}} \tag{10.11}$$

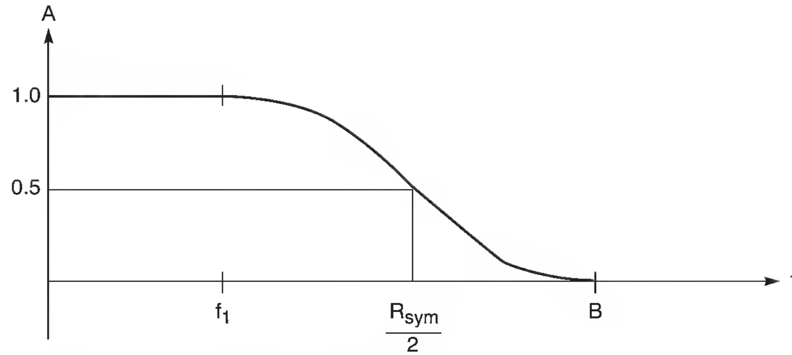


Figure 10.10 The raised cosine response.

and

$$f_1 = \frac{1 - \rho}{2} R_{\text{sym}} \tag{10.12}$$

For binary transmission, the symbol rate simply becomes the bit rate. Thus, for the T1 signal, the required baseband bandwidth is

$$B = \frac{1 + \rho}{2} \times 1.544 \times 10^6$$

$$= 0.772(1 + \rho) \text{ MHz} \tag{10.13}$$

For a rolloff factor of unity, the bandwidth for the T1 system becomes 1.544 MHz.

Although a satellite link requires the use of a modulated carrier wave, the same overall baseband response is needed for the avoidance of ISI. Fortunately, the channel for a satellite link does not introduce frequency distortion, so the pulse shaping can take place in the transmit and receive filters. The modulation of the baseband signal onto a carrier is discussed in the following section.

10.6 Digital Carrier Systems

For transmission to and from a satellite, the baseband digital signal must be modulated onto a microwave carrier. In general, the digital baseband signals may be multilevel (M-ary), requiring multilevel modulation methods. The main binary modulation methods are illustrated in Fig. 10.11. They are defined as follows:

On-off keying (OOK), also known as amplitude-shift keying (ASK). The binary signal in this case is unipolar and is used to switch the carrier on and off.

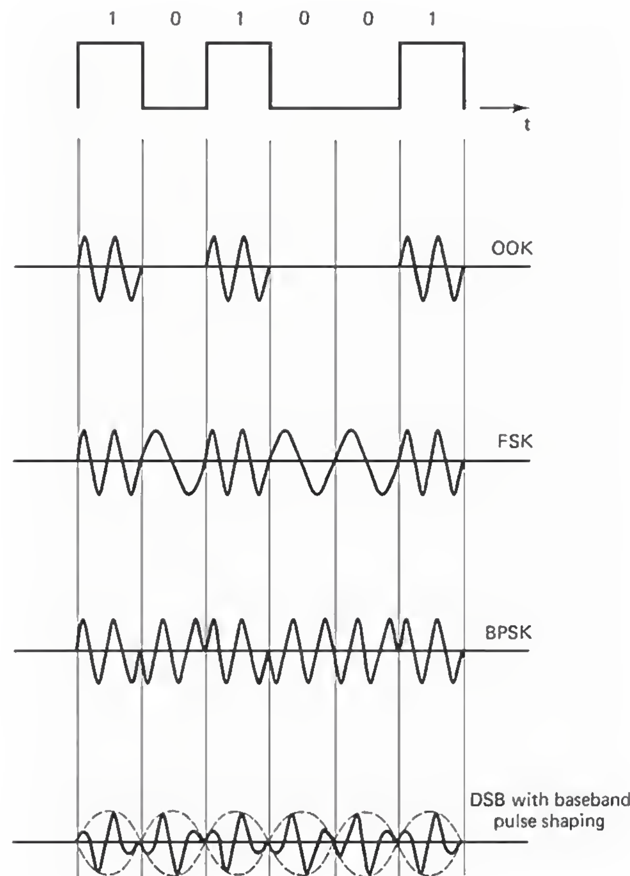


Figure 10.11 Some binary digital modulation formats.

Frequency-shift keying (FSK). The binary signal is used to frequency modulate the carrier, one frequency being used for a binary 1 and another for a binary 0. These are also referred to as the *mark-space frequencies.*

Binary phase-shift keying (BPSK). Polarity changes in the binary signal are used to produce 180° changes in the carrier phase. This may be achieved through the use of double-sideband, suppressed-carrier modulation (DSBSC), with the binary signal as a polar NRZ waveform. In effect, the carrier amplitude is multiplied by a ±1 pulsed waveform. When the binary signal is +1, the carrier sinusoid is unchanged, and when it is −1, the carrier sinusoid is changed in phase by 180°. BPSK is also known as *phase-reversal keying* (PRK). The binary signal may be filtered at baseband before modulation, to

limit the sidebands produced, and as part of the filtering needed for the reduction of ISI, as described in Sec. 10.5. The resulting modulated waveform is sketched in Fig. 10.11.

Differential phase-shift keying (DPSK). This is phase-shift keying in which the phase of the carrier is changed only if the current bit differs from the previous one. A reference bit must be sent at the start of message, but otherwise the method has the advantage of not requiring a reference carrier at the receiver for demodulation.

Quadrature phase-shift keying (QPSK). This is phase-shift keying for a 4-symbol waveform, adjacent phase shifts being equispaced by 90°. The concept can be extended to more than four levels, when it is denoted as MPSK for *M-ary phase-shift keying*.

Quadrature amplitude modulation (QAM). This is also a multilevel (meaning higher than binary) modulation method in which the amplitude and the phase of the carrier are modulated.

Although all the methods mentioned find specific applications in practice, only BPSK and QPSK will be described here, since many of the general properties can be illustrated through these methods, and they are widely used.

10.6.1 Binary phase-shift keying

Binary phase-shift keying may be achieved by using the binary polar NRZ signal to multiply the carrier, as shown in Fig. 10.12*a*. For a binary signal $p(t)$, the modulated wave may be written as

$$e(t) = p(t)\cos\omega_0 t \tag{10.14}$$

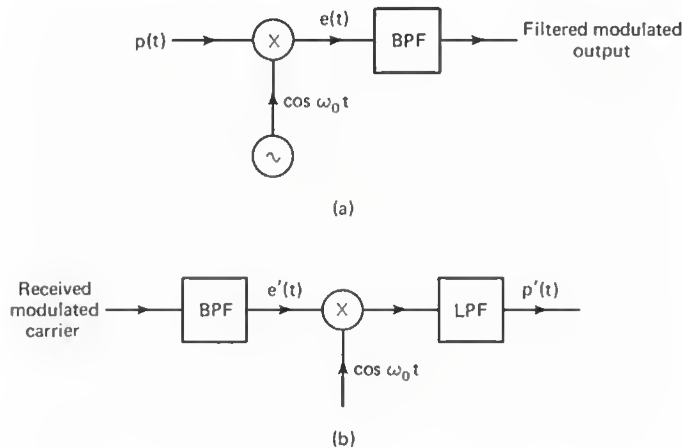


Figure 10.12 (*a*) BPSK modulator; (*b*) coherent detection of a BPSK signal.

When $p(t) = +1$, $e(t) = \cos \omega_0 t$, and when $p(t) = -1$, $e(t) = -\cos \omega_0 t$, which is equivalent to $\cos(\omega_0 t \pm 180°)$. Bandpass filtering of the modulated wave may be used instead of baseband filtering to limit the radiated spectrum. The bandpass filter also may incorporate the square root of the raised-cosine rolloff, described in Sec. 10.5, required to reduce ISI (see, for example, Pratt and Bostian, 1986).

At the receiver (Fig. 10.12*b*), the received modulated carrier will undergo further bandpass filtering to complete the raised-cosine response and to limit input noise. The filtered modulated wave, $e'(t) = p'(t) \cos \omega_0 t$, is passed into another multiplier circuit, where it is multiplied by a replica of the carrier wave $\cos \omega_0 t$. The output from the multiplier is therefore equal to $p'(t) \cos^2 \omega_0 t$. This can be expanded as $p'(t)(0.5 + 0.5 \cos 2\omega_0 t)$. The low-pass filter is used to remove the second harmonic component of the carrier, leaving the low-frequency output, which is $0.5p'(t)$, where $p'(t)$ is the filtered version of the input binary wave $p(t)$. It will be seen that the modulator is basically the same as that used to produce the DSBSC signal described in Sec. 9.3. In the present instance, the bandpass filter following the modulator is used to select the complete DSBSC signal rather than a single sideband.

The receiver is shown in more detail in Fig. 10.13. As shown, a locally generated version of the unmodulated carrier wave is required as one of the inputs to the multiplier. The locally generated carrier has to be exactly in phase with the incoming carrier, and hence this type of detection is termed *coherent detection*. Coherent detection necessitates recovering the unmodulated carrier phase information from the incoming modulated wave, and this is achieved in the *carrier recovery* (CR) section shown in Fig. 10.13.

As discussed in Sec. 10.5, to avoid ISI, sampling must be carried out at the bit rate and at the peaks of the output pulses. This requires the

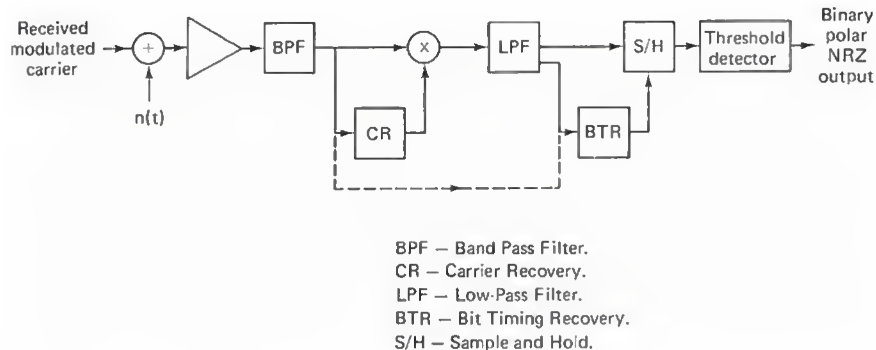


Figure 10.13 Block schematic of a coherent detector showing the carrier recovery section and the bit timing recovery.

sample-and-hold circuit to be accurately synchronized to the bit rate, which necessitates a *bit timing recovery* (BTR) section, as shown in Fig. 10.13.

Thermal noise at the receiver will result in noise phase modulation of the carrier, and so the demodulated waveform $p'(t)$ will be accompanied by noise. The noisy $p'(t)$ signal is passed into the threshold detector which regenerates a noise-free output but one containing some bit errors as a result of the noise already present on the waveform.

The QPSK signal has many features in common with BPSK and will be examined before describing in detail the carrier and bit timing recovery circuits and the effects of noise.

10.6.2 Quadrature phase-shift keying

With QPSK, the binary data are converted into 2-bit symbols which are then used to phase modulate the carrier. Since four combinations containing 2 bits are possible from a binary alphabet (logical 1s and 0s), the carrier phase can be shifted to one of four states.

Figure 10.14*a* shows one way in which QPSK modulation can be achieved. The incoming bit stream $p(t)$ is converted in the serial-to-parallel converter into two binary streams. The conversion is illustrated by the waveforms of Fig. 10.14*b*. For illustration purposes, the bits in the $p(t)$ waveform are labeled *a, b, c, d, e,* and *f*. The serial-to-parallel converter switches bit *a* to the I port and at the same time switches bit *b* to the Q port. In the process, each bit duration is doubled, so the bit rates at the I and Q outputs are half that of the input bit rate.

The $p_i(t)$ bit stream is combined with a carrier $\cos\omega_0 t$ in a BPSK modulator, while the $p_q(t)$ bit stream is combined with a carrier $\sin\omega_0 t$, also in a BPSK modulator. These two BPSK waveforms are added to give the QPSK wave, the various combinations being shown in Table 10.1.

The phase-modulation angles are shown in the phasor diagram of Fig. 10.15. Because the output from the I port modulates the carrier directly, it is termed the *in-phase component*, and hence the designation I. The output from the Q port modulates a quadrature carrier, one which is shifted by 90° from the reference carrier, and hence the designation Q.

Because the modulation is carried out at half the bit rate of the incoming data, the bandwidth required by the QPSK signal is exactly half that required by a BPSK signal carrying the same input data. This is the advantage of QPSK compared with BPSK modulation. The disadvantage is that the modulator and demodulator circuits are more complicated, being equivalent essentially to two BPSK systems in parallel.

Demodulation of the QPSK signal may be carried out by the circuit shown in the block schematic of Fig. 10.16. With the incoming carrier represented as $p_i(t)\cos\omega_0 t - p_q(t)\sin\omega_0 t$, it is easily shown that after

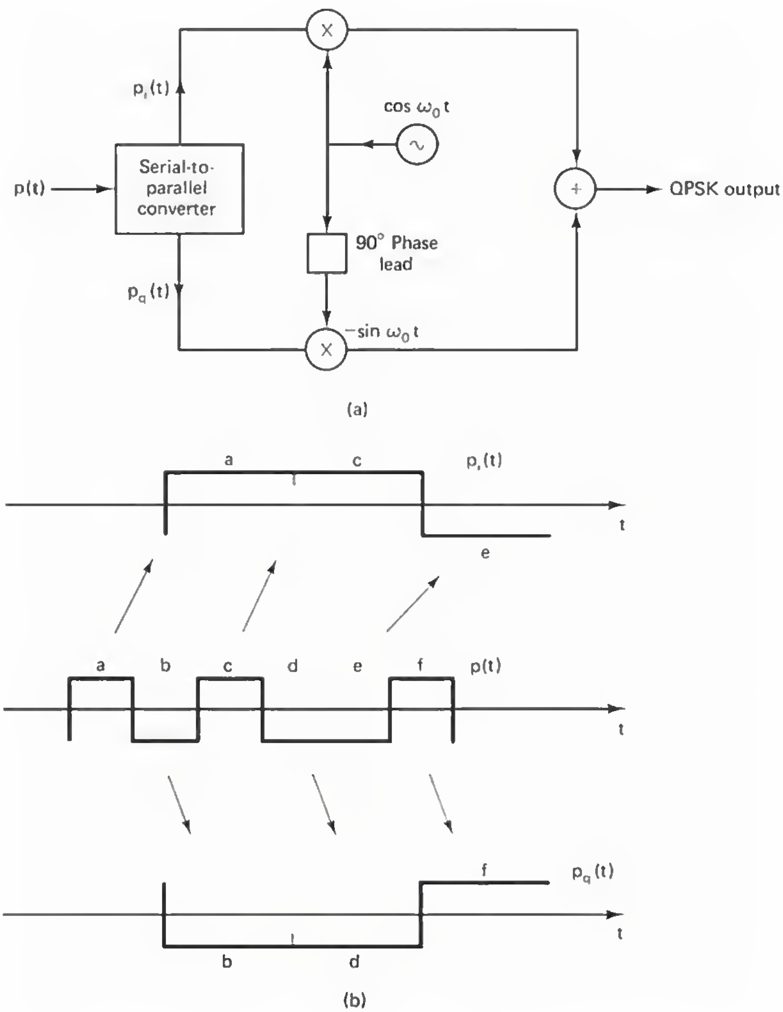


Figure 10.14 (*a*) QPSK modulator; (*b*) waveforms for (*a*).

TABLE 10.1 QPSK Modulator States

$p_i(t)$	$p_q(t)$	QPSK
1	1	$\cos \omega_0 t - \sin \omega_0 t = \sqrt{2} \cos(\omega_0 t + 45°)$
1	−1	$\cos \omega_0 t + \sin \omega_0 t = \sqrt{2} \cos(\omega_0 t - 45°)$
−1	1	$-\cos \omega_0 t - \sin \omega_0 t = \sqrt{2} \cos(\omega_0 t + 135°)$
−1	−1	$-\cos \omega_0 t + \sin \omega_0 t = \sqrt{2} \cos(\omega_0 t - 135°)$

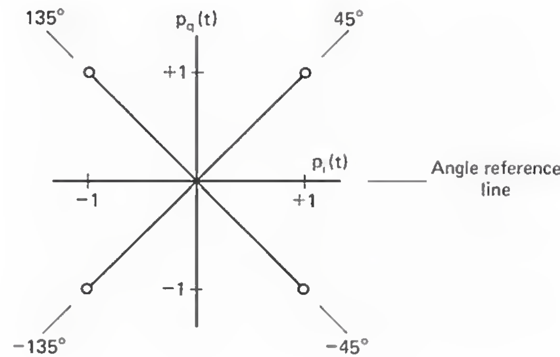


Figure 10.15 Phase diagram for QPSK modulation.

low-pass filtering, the output of the upper BPSK demodulator is $0.5p_i(t)$ and the output of the lower BPSK demodulator is $0.5p_q(t)$. These two signals are combined in the parallel-to-serial converter to yield the desired output $p(t)$. As with the BPSK signal, noise will create errors in the demodulated output of the QPSK signal.

10.6.3 Transmission rate and bandwidth for PSK modulation

Equation (10.14), which shows the baseband signal $p(t)$ multiplied onto the carrier $\cos\omega_0 t$, is equivalent to double-sideband, suppressed-carrier modulation. The digital modulator circuit of Fig. 10.12a is similar to the single-sideband modulator circuit shown in Fig. 9.2, the difference being that after the multiplier, the digital modulator requires a bandpass filter, while the analog modulator requires a single-sideband filter. As

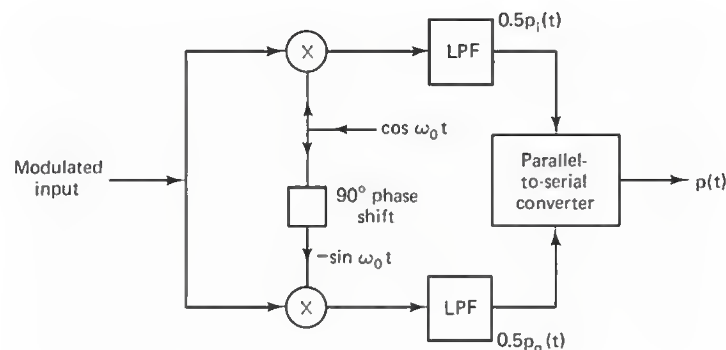


Figure 10.16 Demodulator circuit for QPSK modulation.

shown in Fig. 9.1, the DSBSC spectrum extends to twice the highest frequency in the baseband spectrum. For BPSK modulation the latter is given by Eq. (10.11) with R_{sym} replaced with R_b:

$$B_{IF} = 2B = (1 + \rho)R_b \tag{10.15}$$

Thus, for BPSK with a rolloff factor of unity, the IF bandwidth in hertz is equal to twice the bit rate in bits per second.

As shown in the previous section, QPSK is equivalent to the sum of two orthogonal BPSK carriers, each modulated at a rate $R_b/2$, and therefore, the symbol rate is $R_{sym} = R_b/2$. The spectra of the two BPSK modulated waves overlap exactly, but interference is avoided at the receiver because of the coherent detection using quadrature carriers. Equation (10.15) is modified for QPSK to

$$\begin{aligned} B_{IF} &= (1 + \rho)R_{sym} \\ &= \frac{1 + \rho}{2}R_b \end{aligned} \tag{10.16}$$

An important characteristic of any digital modulation scheme is the ratio of data bit rate to transmission bandwidth. The units for this ratio are usually quoted as bits per second per hertz (a dimensionless ratio in fact because it is equivalent to bits per cycle). Note that it is the data bit rate R_b and not the symbol rate R_{sym} which is used.

For BPSK, Eq. (10.15) gives an R_b/B_{IF} ratio of $1/(1 + \rho)$, and for QPSK, Eq. (10.16) gives an R_b/B_{IF} ratio of $2/(1 + \rho)$. Thus QPSK is twice as efficient as BPSK in this respect. However, more complex equipment is required to generate and detect the QSPK modulated signal.

10.6.4 Bit error rate for PSK modulation

Referring back to Fig. 10.13, the noise at the input to the receiver can cause errors in the detected signal. The noise voltage, which adds to the signal, fluctuates randomly between positive and negative values, and thus the sampled value of signal plus noise may have the opposite polarity to that of the signal alone. This would constitute an error in the received pulse. The noise can be represented by a source at the front of the receiver, shown in Fig. 10.13 (this is discussed in detail in Chap. 12). It is seen that the noise is filtered by the receiver input filter. Thus the receive filter, in addition to contributing to minimizing the ISI, must minimize noise while maximizing the received signal. In short, it must maximize the received signal-to-noise ratio. In practice for satellite links (or radio links), this usually can be

achieved by making the transmit and receive filters identical, each having a frequency response which is the square root of the raised-cosine response. Having identical filters is an advantage from the point of view of manufacturing.

The most commonly encountered type of noise has a flat frequency spectrum, meaning that the noise power spectrum density, measured in joules (or W/Hz), is constant. The noise spectrum density will be denoted by N_0. When the filtering is designed to maximize the received signal-to-noise ratio, the maximum signal-to-noise voltage ratio is found to be equal to $\sqrt{2E_b/N_0}$, where E_b is the average bit energy. The average bit energy can be calculated knowing the average received power P_R and the bit period T_b.

$$E_b = P_R T_b \tag{10.17}$$

The probability of the detector making an error as a result of noise is given by

$$P_e = \frac{1}{2}\text{erfc}\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{10.18}$$

where erfc stands for *complementary error function*, a function whose value is available in tabular or graphic form in books of mathematical tables and as built-in functions in many computational packages. A related function, called the *error function*, denoted by *erf*(·) is sometimes used, where

$$\text{erfc}(x) = 1 - \text{erf}(x) \tag{10.19}$$

Equation (10.18) applies for polar NRZ baseband signals and for BPSK and QPSK modulation systems. The probability of bit error is also referred to as the *bit error rate* (BER). A P_e of 10^{-6} signifies a BER of 1 bit in a million, on average. The graph of P_e versus E_b/N_0 in decibels is shown in Fig. 10.17. Note carefully that the energy ratio, not the decibel value, of E_b/N_0 must be used in Eq. (10.18). This is illustrated in the following example.

Example 10.1 The average power received in a binary polar transmission is 10 mW, and the bit period is 100 μs. If the noise power spectral density is 0.1 μJ, and optimum filtering is used, determine the bit error rate.

Solution From Eq. (10.17):

$$\begin{aligned} E_b &= 10 \times 10^{-3} \times 100 \times 10^{-6} \\ &= 10^{-6}\text{J} \end{aligned}$$

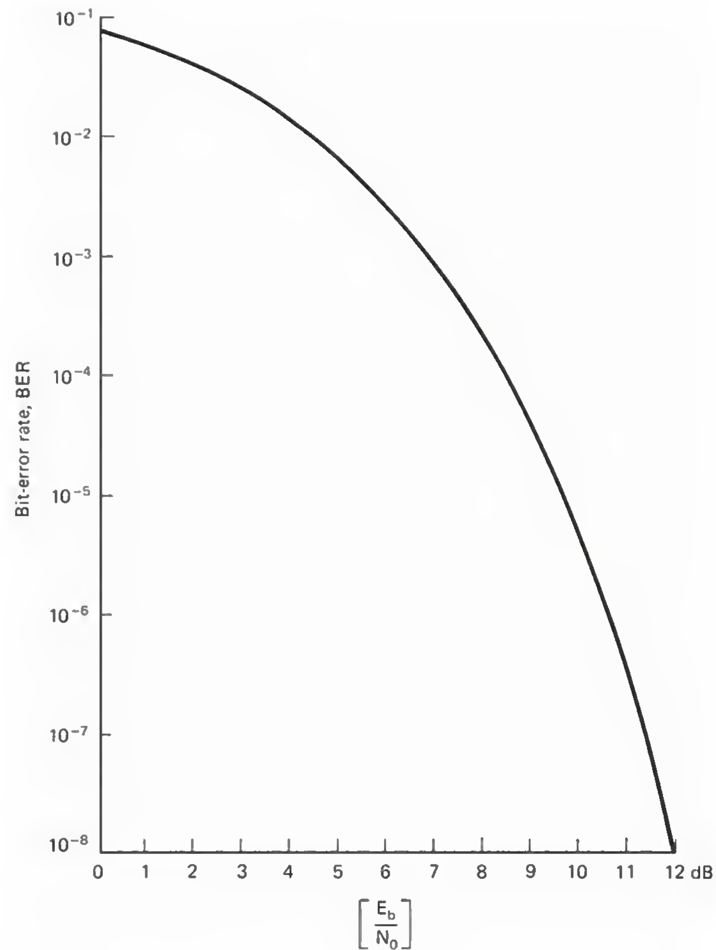


Figure 10.17 BER versus (E_b/N_0) for baseband signaling using a binary polar NRZ waveform. The curve also applies for BPSK and QPSK modulated waveforms.

and

$$\frac{E_b}{N_0} = \frac{10^{-6}}{10^{-7}}$$
$$= 10$$

$$\text{erf}(\sqrt{10}) \cong 0.9999923$$

Combining Eqs. (10.18) and (10.19):

$$\text{BER} = 0.5(1 - 0.9999923)$$
$$= \underline{\underline{3.9 \times 10^{-6}}}$$

Equation (10.18) is sometimes expressed in the alternative form

$$P_e = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \tag{10.20}$$

Here, the $Q(\cdot)$ function is simply an alternative way of expressing the complementary error function, and in general

$$\text{erfc}(x) = 2Q(\sqrt{2}x) \tag{10.21}$$

These relationships are given for reference only and will not be used further in this book.

An important parameter for carrier systems is the ratio of the average carrier power to the noise power density, usually denoted by $[C/N_0]$. The $[E_b/N_0]$ and $[C/N_0]$ ratios can be related as follows. The average carrier power at the receiver is P_R W. The energy per symbol is therefore P_R/R_{sym} J, with R_{sym} in symbols per second. Since each symbol contains m bits, the energy per bit is P_R/mR_{sym} J. But $mR_{\text{sym}} = R_b$, and therefore, the energy per bit, E_b, is

$$E_b = \frac{P_R}{R_b} \tag{10.22}$$

As before, let N_0 represent the noise power density. Then $E_b/N_0 = P_R/R_bN_0$. But P_R/N_0 is the carrier-to-noise density ratio, usually denoted by C/N_0, and therefore,

$$\frac{E_b}{N_0} = \frac{C/N_0}{R_b} \tag{10.23}$$

Rearranging this and putting it in decibel notation gives

$$\left[\frac{C}{N_0}\right] = \left[\frac{E_b}{N_0}\right] + [R_b] \tag{10.24}$$

It should be noted that whereas $[E_b/N_0]$ has units of decibels, $[C/N_0]$ has units of dBHz, as explained in App. G.

Example 10.2 The downlink transmission rate in a satellite circuit is 61 Mb/s, and the required $[E_b/N_0]$ at the ground station receiver is 9.5 dB. Calculate the required $[C/N_0]$.

Solution The transmission rate in decibels is $[R_b] = 10\log(61 \times 10^6) = 77.85$ dBb/s

Hence

$$\left[\frac{C}{N_0}\right] = 77.85 + 9.5 = \underline{\underline{87.35 \text{ dBHz}}}$$

The equations giving the probability of bit error are derived on the basis that the filtering provides maximum signal-to-noise ratio. In practice, there are a number of reasons why the optimal filtering may not be achieved. The raised-cosine response is a theoretical model that can only be approximated in practice. Also, for economic reasons, it is desirable to use production filters manufactured to the same specifications for the transmit and receive filter functions, and this may result in some deviation from the desired theoretical response. The usual approach in practice is that one knows the BER that is acceptable for a given application. The corresponding ratio of bit energy to noise density can then be found from Eq. (10.18) or from a graph such as that shown in Fig. 10.17. Once the theoretical value of E_b/N_0 is found, an *implementation margin*, amounting to a few decibels at most, is added to allow for imperfections in the filtering. This is illustrated in the following example.

Example 10.3 A BPSK satellite digital link is required to operate with a bit error rate of no more than 10^{-5}, the implementation margin being 2 dB. Calculate the required E_b/N_0 ratio in decibels.

Solution The graph of Fig. 10.17 shows that E_b/N_0 is around 9 dB for a BER of 10^{-5}. By plotting this region to an expanded scale, a more accurate value of E_b/N_0 can be obtained. This is shown in Fig. 10.18. from which $[E_b/N_0]$ is seen to be about 9.65 dB. This is without an implementation margin. The required value, including an implementation margin, is $9.65 + 2 = \underline{\underline{11.65 \text{ dB}}}$.

To summarize, BER is a specified requirement, which enables E_b/N_0 to be determined by using Eq. (10.18) or Fig. 10.17. The rate R_b also will be specified, and hence the $[C/N_0]$ ratio can be found by using Eq. (10.24).

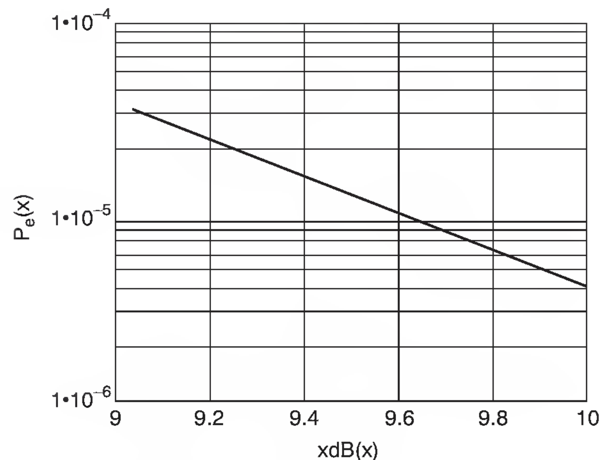
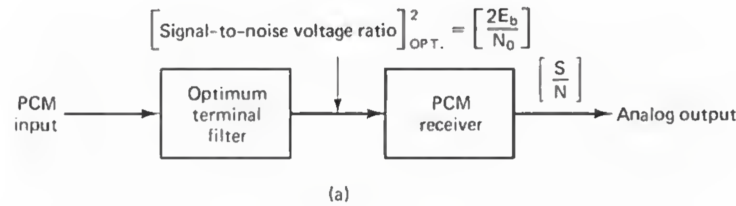


Figure 10.18 Solution for Example 10.3.

The $[C/N_0]$ ratio is then used in the link budget calculations, as described in Chap. 12.

With purely digital systems, the BER will be directly reflected in errors in the data being transmitted. With analog signals which have been converted to digital form through PCM, the BER contributes to the output signal-to-noise ratio, along with the quantization noise, as described in Sec. 10.3. Curves showing the contributions of thermal noise and quantization noise to the signal-to-noise output for analog systems can be seen in Fig. 10.19. The signal-to-noise power ratio is given by (Taub and Schilling, 1986)

$$\frac{S}{N} = \frac{Q^2}{1 + 4Q^2 P_e} \tag{10.25}$$



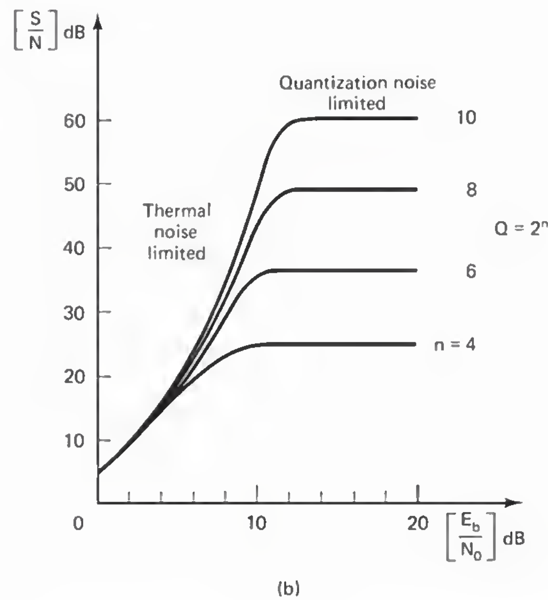


Figure 10.19 (*a*) Use of optimum terminal filter to maximize the signal-to-noise voltage ratio; (*b*) plot of Eq. (10.25).

where $Q = 2^n$ is the number of quantized steps, and n is the number of bits per sample.

The BER can be improved through the use of error control coding. This is the topic of Chap. 11.

10.7 Carrier Recovery Circuits

To implement coherent detection, a *local oscillator* (LO) that is exactly synchronized to the carrier must be provided at the receiver. As shown in Sec. 10.6.1, a BPSK signal is a *double sideband suppressed carrier* (DSBSC) type of signal, and therefore, the carrier is not directly available in the BPSK signal. The carrier can be recovered using a *squaring loop*, as shown in Fig. 10.20. Consider first the situation where the input is a BPSK signal. The frequency multiplier is a nonlinear circuit, which squares the signal. Squaring Eq. (10.14) results in

$$\begin{aligned} e^2(t) &= p^2(t)\cos^2\omega_0 t \\ &= p^2(t)\left(\frac{1}{2} + \frac{1}{2}\cos 2\omega_0 t\right) \end{aligned} \tag{10.26}$$

Note that with $p(t)$ equal to ± 1, the square is just 1. The bandpass filter following the frequency multiplier is tuned to the carrier second harmonic, which provides one of the inputs to the phase detector of the phase-locked loop. The *voltage-controlled oscillator* (VCO) in the *phase-locked loop* (PLL) operates at the carrier frequency. The second frequency multiplier provides the second harmonic of this as the other input to the phase detector. The phase difference between these two inputs generates a bias voltage that brings the frequency of the VCO into synchronism with the carrier frequency as derived from the BPSK signal.

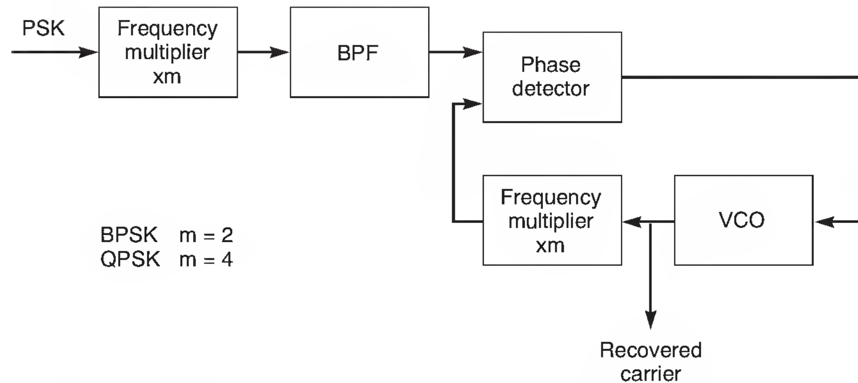


Figure 10.20 Functional block diagram for carrier recovery.

With QPSK, the signal can be represented by the formulas given in Table 10.1, which may be written generally as

$$e(t) = \sqrt{2}\cos\left(\omega_0 t \pm \frac{n\pi}{4}\right) \quad (10.27)$$

Quadrupling this, followed by some trigonometric simplification, results in

$$e^4(t) = \frac{3}{2} + 2\cos 2\left(\omega_0 t \pm \frac{n\pi}{4}\right) + \frac{1}{2}\cos 4\left(\omega_0 t \pm \frac{n\pi}{4}\right) \quad (10.28)$$

The last term on the right-hand side is selected by the bandpass filter and is

$$\frac{1}{2}\cos 4\left(\omega_0 t \pm \frac{n\pi}{4}\right) = \frac{1}{2}\cos(4\omega_0 t \pm n\pi) \quad (10.29)$$

This is seen to consist of the fourth harmonic of the carrier, including a constant-phase term that can be ignored. The fourth harmonic is selected by the bandpass filter, and the operation of the circuit proceeds in a similar manner to that for the BPSK signal.

Frequency multiplication can be avoided by use of a method known as the *Costas loop*. Details of this, along with an analysis of the effects of noise on the squaring loop and the Costas loop methods, will be found in Gagliardi (1991). Other methods are also described in detail in Franks (1980).

10.8 Bit Timing Recovery

Accurate bit timing is needed at the receiver in order to be able to sample the received waveform at the optimal points. In the most common arrangements, the clocking signal is recovered from the demodulated waveform, these being known as *self-clocking* or *self-synchronizing systems*. Where the waveform has a high density of zero crossings, a zero-crossing detector can be used to recover the clocking signal. In practice, the received waveform is often badly distorted by the frequency response of the transmission link and by noise, and the design of the bit timing recovery circuit is quite complicated. In most instances, the spectrum of the received waveform will not contain a discrete component at the clock frequency. However, it can be shown that a periodic component at the clocking frequency is present in the squared waveform for digital signals (unless the received pulses are exactly rectangular, in which case squaring simply produces a dc level for a binary waveform). A commonly used baseband scheme is

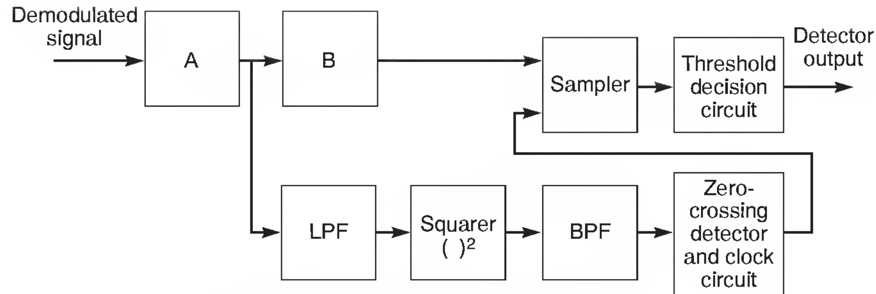


Figure 10.21 Functional block diagram for bit-timing recovery.

shown in block schematic form in Fig. 10.21 (Franks, 1980). The filters A and B form part of the normal signal filtering (e.g., raised-cosine filtering). The signal for the bit timing recovery is tapped from the junction between A and B and passed along a separate branch which consists of a filter, a squaring circuit, and a bandpass filter which is sharply tuned to the clock frequency component present in the spectrum of the squared signal. This is then used to synchronize the clocking circuit, the output of which clocks the sampler in the detector branch.

The *early-late gate circuit* provides a method of recovering bit timing which does not rely on a clocking component in the spectrum of the received waveform. The circuit utilizes a feedback loop in which the magnitude changes in the outputs from matched filters control the frequency of a local clocking circuit (for an elementary description see, for example, Roddy and Coolen, 1995). Detailed analyses of these and other methods will be found in Franks (1980) and Gagliardi (1991).

10.9 Problems and Exercises

10.1. For a test pattern consisting of alternating binary 1s and 0s, determine the frequency spectra in terms of the bit period T_b for the following signal formats: (*a*) unipolar; (*b*) polar NRZ; (*c*) polar RZ; (*d*) Manchester.

10.2. Plot the raised-cosine frequency response Eq. (10.9), for a bit rate of 1 b/s and a roll of factor of 1, for a symbol rate equal to the bit rate. Use the inverse Fourier transform to determine the shape of the pulse time waveform.

10.3. Plot the compressor transfer characteristics for $\mu = 100$ and $A = 100$. The μ-law compression characteristic is given by

$$v_o = sign(v_i)\frac{\ln(1 + \mu|v_i|)}{\ln(1 + \mu)}$$

where v_0 is the output voltage normalized to the maximum output voltage, and v_i is the input voltage normalized to the maximum input voltage. The A-law characteristic is given by

$$v_0 = sign(v_i)\frac{A|v_i|}{1 + \ln A} \quad \text{for} \quad 0 \leq |v_i| \leq \frac{1}{A}$$

and

$$v_0 = sign(v_i)\frac{1 + \ln(A|v_i|)}{1 + \ln A} \quad \text{for} \quad \frac{1}{A} \leq |v_i| \leq 1$$

10.4. Write down the expander transfer characteristics corresponding to the compressor characteristics given in Prob. 10.3.

10.5. Assuming that the normalized levels shown in Fig. 10.4*b* represent millivolts, write out the digitally encoded words for input levels of (*a*) ±90 mV; (*b*) ±100 mV; (*c*) ±190 mV; (*d*) ±3000 mV.

10.6. Determine the decoded output voltage levels for the input levels given in Prob. 10.5. Determine also the quantization error in each case.

10.7. (*a*) A test tone having the full peak-to-peak range is applied to a PCM system. If the number of bits per sample is 8, determine the quantization *S*/*N*. Assume uniform sampling of step size ΔV, for which the mean square noise voltage is $(\Delta V)^2/12$. (*b*) Given that a raised-cosine filter is used with $\rho = 1$, determine the bandwidth expansion factor B/W, where B is the PCM bandwidth and W is the upper cutoff frequency of the input.

10.8. A PCM signal uses the polar NRZ format. Following optimal filtering, the $[E_b/N_0]$ at the input to the receiver decision detector is 10 dB. Determine the bit error rate (BER) at the output of the decision detector.

10.9. Using Eq. (10.18), calculate the probability of bit error for $[E_b/N_0]$ values of (*a*) 0 dB, (*b*) 10 dB, and (*c*) 40 dB.

10.10. A PCM system uses 8 bits per sample and polar NRZ transmission. Determine the output $[S/N]$ for $[E_b/N_0]$ values of (*a*) 0 dB, (*b*) 10 dB, and (*c*) 40 dB at the input to the decision detector.

10.11. A binary periodic waveform of period $3T_b$ is low-pass filtered before being applied to a BPSK modulator. The low-pass filter cuts off at $B = 0.5/T_b$. Derive the trigonometric expansion for the modulated wave, showing that only side frequencies and no carrier are present. Given that the bit period is 100 ms and the carrier frequency is 100 kHz, sketch the spectrum, showing the frequencies to scale.

10.12. Explain what is meant by *coherent detection* as used for the demodulation of PSK bandpass signals. An envelope detector is an example of a *noncoherent detector*. Can such a detector be used for BPSK? Give reasons for your answer.

10.13. Explain how a QPSK signal can be represented by two BPSK signals. Show that the bandwidth required for QPSK signal is one-half that required for a BPSK signal operating at the same data rate.

10.14. The input data rate on a satellite circuit is 1.544 Mbps. Calculate the bandwidths required for BPSK modulation and for QPSK modulation, given that raised-cosine filtering is used with a rolloff factor of 0.2 in each case.

10.15. A QPSK system operates at a $[E_b/N_0]$ ratio of 8 dB. Determine the bit error rate.

10.16. A BPSK system operates at a $[E_b/N_0]$ ratio of 16 dB. Determine the bit error rate.

10.17. The received power in a satellite digital communications link is 0.5 pW. The carrier is BPSK modulated at a bit rate of 1.544 Mb/s. If the noise power density at the receiver is 0.5×10^{-19} J, determine the bit error rate.

10.18. The received $[C/N_0]$ ratio in a digital satellite communications link is 86.5 dBHz, and the data bit rate is 50 Mb/s. Calculate the $[E_b/N_0]$ ratio and the BER for the link.

10.19. For the link specified in Prob. 10.18, the $[C/N_0]$ ratio is improved to 87.5 dBHz. Determine the new BER.

References

Bellamy, J. 1982. *Digital Telephony*. Wiley, New York.

Franks, L. E. 1980. "Carrier and Bit Synchronization in Data Communication: A Tutorial Review." *IEEE Trans. Commun.*, Vol. 28, No. 8, August, pp. 1107–1120.

Gagliardi, R. M. 1991. *Satellite Communications,* 2d ed. Van Nostrand Reinhold, New York.

Hassanein, H., A. B. Amour, and K. Bryden. 1992. "A Hybrid Multiband Excitation Coder for Low Bit Rates." Department of Communications, Communications Research Centre, Ottawa, Ontario, Canada.

Hassanein, H., A. B. Amour, K. Bryden, and R. Deguire. 1989. "Implementation of a 4800 bps Code-Excited Linear Predictive Coder on a Single TMS320C25 Chip." Department of Communications, Communications Research Centre, Ottawa, Ontario, Canada.

Pratt, T., and C. W. Bostian. 1986. *Satellite Communications*. Wiley, New York.

Roddy, D., and J. Coolen. 1994. *Electronic Communications,* 4th ed. Prentice-Hall, Englewood Cliffs, NJ.

Taub, H., and D. L. Schilling. 1986. *Principles of Communications Systems,* 2d ed. McGraw-Hill, New York.

Chapter

11

Error Control Coding

11.1 Introduction

As shown by Fig. 10.17, the probability of bit error (P_e) in a digital transmission can be reduced by increasing $[E_b/N_0]$, but there are practical limits to this approach. Equation (10.24) shows that for a given bit rate R_b, $[E_b/N_0]$ is directly proportional to $[C/N_0]$. An increase in $[C/N_0]$ can be achieved by increasing transmitted power and/or reducing the system noise temperature (to reduce N_0). Both these measures are limited by cost and, in the case of the onboard satellite equipment, size. In practical terms, a probability of bit error (P_e of Eq. 10.18) of about 10^{-4}, which is satisfactory for voice transmissions, can be achieved with off the-shelf equipment. For lower P_e values such as required for some data, error control coding must be used. Error control performs two functions, error detection and error correction. Most codes can perform both functions, but not necessarily together. In general, a code is capable of detecting more errors than it can correct. Where error detection only is employed, the receiver can request a repeat transmission (a technique referred to as *automatic repeat request*, or ARQ). This is only of limited use in satellite communications because of the long transmission delay time associated with geostationary satellites, and of course radio and TV broadcast is essentially one-way so ARQ cannot be employed. What is termed *forward error correction* (FEC) allows errors to be corrected without the need for retransmission, but this is more difficult and costly to implement than ARQ.

A P_e value of 10^{-4} represents an average error rate of 1 bit in 10^4, and the error performance is sometimes specified as the *bit error rate* (BER). It should be recognized, however, that the probability of bit error P_e occurs as a result of noise at the input to the receiver, while the BER is the actual error rate at the output of the detector. When error control

coding is employed, the distinction between P_e and BER becomes important. P_e is still determined by conditions at the input, but the error control will, if properly implemented, make the probability of bit error at the output (the BER) less than that at the input. Error control coding applies only to digital signals, and in most cases the signal is in binary form, where the message symbols are bits, or logic 1s and 0s.

Encoding refers to the process of adding coding bits to the uncoded bit stream, and *decoding* refers to the process of recovering the original (uncoded) bit stream from the coded bit stream. Both processes are usually combined in one unit termed a *codec*.

11.2 Linear Block Codes

A block code requires that the message stream be partitioned into blocks of bits (considering only binary messages at this stage). Let each block contain k bits, and let these k bits define a dataword. The number of datawords is 2^k. There is no redundancy in the system, meaning that even a single bit error in transmission would convert one dataword into another, which of course would constitute an error.

The datawords can be encoded into codewords which consist of n bits, where the additional $n-k$ bits are derived from the message bits but are not part of the message. The number of possible codewords is 2^n, but only 2^k of these will contain datawords, and these are the ones that are transmitted. It follows that the rest of the codewords are redundant, but only in the sense that they do not contribute to the message. (The $n-k$ additional bits are referred to as *parity check bits*). If errors occur in transmission, there is high probability that they will convert the permissible codewords into one or another of the redundant words that the decoder at the receiver is designed to recognize as an error. It will be noted that the term *high probability* is used. There is always the possibility, however remote, that enough errors occur to transform a transmitted codeword into another legitimate codeword in error.

The code *rate* r_c is defined as the ratio of dataword bits to codeword bits (note that although it is called a *rate*, it is not a rate in bits per second)

$$r_c = \frac{k}{n} \tag{11.1}$$

The code is denoted by (n, k) for example a code which converts a 4-bit dataword into a 7 bit codeword would be a (7, 4) code.

A repetition code illustrates some of the general properties of block codes. In a repetition code, each bit is considered to be a dataword, in effect, $k = 1$. For n-redundancy encoding, the output of the encoder is n bits, identical to the input bit. As an example, consider the situation

when $n = 3$. A binary 1 at the input to the encoder results in a 111 codeword at the output, and a binary 0 at the input results in a 000 codeword at the output. At the receiver, the logic circuits in the decoder produce a 1 when 111 is present at the input and a 0 when 000 is present. It is assumed that synchronization is maintained between encoder and decoder. If a codeword other than 111 or 000 is received, the decoder detects an error and can request a retransmission (ARQ).

FEC can take place on the basis of a "majority vote." In this case, the logic circuits in the decoder are designed to produce a 1 at the output whenever two or three 1s occur in the received codeword (codewords 111, 101, 011, and 110) and a 0 whenever two or three 0s appear in the codeword (codewords 000, 001, 010, and 100). An odd number of "repeats" is used to avoid a tied vote.

Errors can still get through if the noise results in two or three successive errors in a codeword. For example, if the noise changes a 111 into a 000 or a 000 into a 111, the output will be in error whether error detection or FEC is used. If two errors occur in a codeword, then the "majority vote" method for FEC will result in an error. However, the probability of two or three consecutive errors occurring is very much less than the probability of a single error. This assumes that the bit energy stays the same, an aspect that is discussed in Sec. 11.7. Codes that are more efficient than repetitive encoding are generally used in practice.

As a matter of definition, a code is termed *linear* when any linear combination of two codewords is also a codeword. For binary codewords in particular, the linear operation encountered is modulo-2 addition. The rules for modulo-2 addition are:

$$0+0=0 \qquad 0+1=1 \qquad 1+0=1 \qquad 1+1=0$$

Modulo-2 addition is easily implemented in hardware using the exclusive-OR (XOR) digital circuit, which is one of its main advantages. All codes encountered in practice are linear, which has a bearing on the theoretical development (see Proakis and Salehi, 1994).

To illustrate this further consider the eight possible datawords formed from a 3-bit sequence. One parity bit will be attached to each dataword as shown in the Table 11.1. The parity bits are selected to provide *even parity*, that is, the number of 1s in any codeword is even (including the all-zero codeword). The parity bits are found by performing a modulo-2 addition on the dataword.

With even parity it is seen that the bits in the codeword modulo-2 sum to zero. It would be possible to use *odd parity* where the parity bit is chosen so that the modulo-2 sum of the codeword is 1, however this would exclude the all-zero codeword. A linear code must include the all-zero codeword, and hence even parity is used.

TABLE 11.1 Even Parity Codewords

Dataword	Modulo-2 addition of dataword	Codeword
000	0	0000
001	1	0011
010	1	0101
011	0	0110
100	1	1001
101	0	1010
110	0	1100
111	1	1111

The *Hamming distance* between two codewords is defined as the number of positions by which the two codewords differ. Thus the codewords 0000 and 1111 differ in four positions, so their Hamming distance is four. The *minimum Hamming distance*, usually just referred to as the *minimum distance* is the smallest Hamming distance between any two codewords. It can be shown that the minimum distance is given by the minimum number of binary 1s in any codeword, excluding the all-zero codeword. By inspection it will be seen that the minimum distance of the code in Table 11.1 is two. The greater the minimum distance the better the code, as this reduces the chances of one codeword being converted to another by noise.

The properties of linear block codes are best formulated in terms of matrices. Only a summary of some of these results are presented here, as background to aid in the understanding of coding methods used in satellite communications. A dataword (or message block) of size k is denoted by a row vector $\mathbf{d}$, for example the sixth dataword in Table 11.1 is $\mathbf{d}_6 = [101]$. Denoting the codeword by row vector $\mathbf{c}$, the corresponding codeword is $\mathbf{c}_6 = [1010]$. In general, the codeword is generated from the dataword by use of a *generator matrix* denoted by $\mathbf{G}$, where

$$\mathbf{c} = \mathbf{dG} \tag{11.2}$$

Design of the generator matrix forms part of coding practice and will not be gone into here. However an example will illustrate the properties. One example of a generator matrix for a (7, 4) code is

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \tag{11.3}$$

It will be noted that the matrix has 7 columns and 4 rows corresponding to the (7, 4) code, and furthermore, the first four columns form an *identity submatrix.* The identity submatrix results in the dataword appearing as the first four bits of the codeword, in this example. In general, a *systematic code* contains a sequence that is the dataword, and the most common arrangement is to have the dataword at the start of the codeword as shown in the example. It can be shown that any linear block code can be put into systematic form. The remaining bits in any row of **G** are responsible for generating the parity bits from the data bits. As an example, suppose it is required to generate a codeword for a dataword [1010]. This is done by multiplying **d** by **G**

$$\mathbf{C} = [1 \quad 0 \quad 1 \quad 0]\begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

$$= [1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0]$$

The dataword is seen to appear as the first four bits in the codeword, and the end three bits are the parity bits. The parity bits are generated from the data bits by means of the last three columns in the generator matrix. This submatrix is denoted by **P**:

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \tag{11.4}$$

The *transpose* of **P**, which enters into the decoding process is formed by interchanging rows with columns, that is, row 1 becomes column 1, and column 1 becomes row 1, row 2 becomes column 2 and column 2 becomes row 2, and so on. In full, the transpose of **P**, written as $\mathbf{P}^{\mathrm{T}}$ is

$$\mathbf{P}^{\mathrm{T}} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

What is termed the *parity check matrix* (denoted by **H**) is now formed by appending an identity matrix to $\mathbf{P}^{\mathrm{T}}$:

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{11.5}$$

The number of rows in $\mathbf{H}$ is equal to the number of parity bits, $n-k$, and the number of columns is n, that is the parity check matrix is a $(n-k, n)$ matrix. A fundamental property of these code matrices is that

$$\mathbf{GH}^{\mathrm{T}} = \mathbf{0} \tag{11.6}$$

When a codeword is received it can be verified as being correct on multiplying it by $\mathbf{H}^{\mathrm{T}}$. The product $\mathbf{cH}^{\mathrm{T}}$ should be equal to zero. This follows since $\mathbf{c} = \mathbf{dG}$ and therefore $\mathbf{cH}^{\mathrm{T}} = \mathbf{dGH}^{\mathrm{T}} = \mathbf{0}$. If a result other than zero is obtained, then an error has been detected. In general terms, the product $\mathbf{cH}^{\mathrm{T}}$ gives what is termed the *syndrome* and denoting this by $\mathbf{s}$:

$$\mathbf{s} = \mathbf{cH}^{\mathrm{T}} \tag{11.7}$$

A received codeword can be represented by the transmitted codeword plus a possible error vector. For example if the transmitted codeword is [1010010] and the received codeword is [1010110] the error is in the fifth bit position from the left and this can be written as

$$[1010110] = [1010010] + [0000100]$$

More generally, if $\mathbf{c}_{\mathrm{R}}$ is the received codeword, $\mathbf{c}_{\mathrm{T}}$ the transmitted codeword and $\mathbf{e}$ the error vector then, with modulo-2 addition

$$\mathbf{c}_{\mathrm{R}} = \mathbf{c}_{\mathrm{T}} + \mathbf{e} \tag{11.8}$$

Substituting $\mathbf{c}_{\mathrm{R}}$ for $\mathbf{c}$ in Eq. (11.7) gives

$$\begin{aligned} \mathbf{s} &= (\mathbf{c}_{\mathrm{T}} + \mathbf{e})\mathbf{H}^{\mathrm{T}} \\ &= \mathbf{c}_{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{eH}^{\mathrm{T}} \end{aligned}$$

But as shown earlier, the product $\mathbf{cH}^{\mathrm{T}}$, which is $\mathbf{c}_{\mathrm{T}}\mathbf{H}^{\mathrm{T}}$ in this notation, is zero, hence,

$$\mathbf{s} = \mathbf{eH}^{\mathrm{T}} \tag{11.9}$$

This shows that the syndrome depends only on the error vector and is independent of the codeword transmitted. Since the error vector has the same number of bits n as the codeword there will be 2^n possible error vectors. Not all of these can be detected since the syndrome has only $n-k$ bits (determined by the number of rows in the $\mathbf{H}$ matrix), giving as 2^{n-k} the number of different syndromes. One of these will be the all zero syndrome, and hence the number of errors that can be detected is just $2^{n-k} - 1$. In practice the decoder is designed to correct the most likely errors, for example those with only 1-bit error. The received syndrome may be compared with

values in a lookup table, (tabulated against known error patterns), and the most likely match found. This is termed maximum likelihood decoding.

The Hamming codes described in the next section enables a single error to be corrected, and in fact the syndrome gives the bit position of the error. Suppose for example the codeword [1010010] as previously determined is transmitted but a bit error occurs in the fifth bit from the left, so that the received codeword is [1010110]. Applying Eq. (11.7)

$$s = [1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 0] \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= [1 \quad 0 \quad 0]$$

The fact that the syndrome is not zero indicates that an error has occurred. For the case of Hamming codes discussed in Sec. 11.3, and of which this is an example, the syndrome also shows which bit is in error. The syndrome [100] corresponds to the fifth column (from the left) of the parity check matrix **H** and this indicates that it is the fifth bit that is in error. In general, with the Hamming code, if the syndrome corresponds to column m then bit m is the one in error, and this can be "flipped" to the correct value.

11.3 Cyclic Codes

Cyclic codes are a subclass of linear block codes. They possess the property that a cyclic shift of a codeword is also a codeword. For example, if a codeword consists of the elements $\{c_1\ c_2\ c_3\ c_4\ c_5\ c_6\ c_7\}$, then $\{c_2\ c_3\ c_4\ c_5\ c_6\ c_7\ c_1\}$ is also a codeword. The advantage of cyclic codes is that they are easily implemented in practice through the use of shift registers and modulo-2 adders. Cyclic codes are widely used in satellite transmission, and the properties of the most important of these are summarized in the following sections. Only certain combinations of k and n are permitted in these codes. As pointed out in Taub and Schilling (1986), a code is an invention, and these codes are named after their inventors.

11.3.1 Hamming codes

For an integer $m \geq 2$, the k and n values are related as $n = 2^m - 1$ and $k = n - m$. Thus some of the permissible combinations are shown in Table 11.2:

TABLE 11.2 *m, n, k* for some Hamming Codes

m	n	k
2	3	1
3	7	4
4	15	11
5	31	26
6	63	57
7	127	120

It will be seen that the code rate $r_c = k/n$ approaches unity as m increases, which leads to more efficient coding. However, only a single error can be corrected with Hamming codes.

11.3.2 BCH codes

BCH stands for the names of the inventors, Bose, Chaudhuri, and Hocquenghen. These codes can correct up to t errors, and where m is any positive integer, the permissible values are $n = 2^m - 1$ and $k \geq n - mt$. The integers m and t are arbitrary, which gives the code designer considerable flexibility in choice. Proakis and Salehi (1994) give an extensive listing of the parameters for BCH codes, from which the values in Table 11.3 have been obtained. As usual, the code rate is $r_c = k/n$.

11.3.3 Reed-Solomon codes

The codes described so far work well with errors that occur randomly rather than in bursts. However, there are situations where errors do occur in bursts; that is, a number of bits that are close together may

Table 11.3 Some parameters for BCH codes

n	k	t
7	4	1
15	11	1
15	7	2
15	5	3
31	26	1
31	21	2
31	16	3
31	11	5
31	6	7

experience errors as a result of impulse-type noise or impulse-type interference. *Reed-Solomon* (R-S) codes are designed to correct errors under these conditions. Instead of encoding directly in bits, the bits are grouped into symbols, and the datawords and codewords are encoded in these symbols. Errors affecting a group of bits are most likely to affect only one symbol that can be corrected by the R-S code.

Let the number of bits per symbol be k; then the number of possible symbols is $q = 2^k$ (referred to as a *q-ary alphabet*). Let K be the number of symbols in a dataword and N be the number of symbols in a codeword. Just as in the block code where k-bit datawords were mapped into n-bit codewords, with the R-S code, datawords of K symbols are mapped into codewords of N symbols. The additional $N - K$ redundant symbols are derived from the message symbols but are not part of the message. The number of possible codewords is 2^N, but only 2^K of these will contain datawords, and these are the ones that are transmitted. It follows that the rest of the codewords are redundant, but only in the sense that they do not contribute to the message. If errors occur in transmission, there is high probability that they will convert the permissible codewords into one or another of the redundant words that the decoder at the receiver is designed to recognize as an error. It will be noted that the term *high probability* is used. There is always the possibility, however remote, that enough errors occur to transform a transmitted codeword into another legitimate codeword even though this was not the one transmitted.

It will be observed that the wording of the preceding paragraph parallels that given in Sec. 11.2 on block codes, except that here the coding is carried out on symbols. Some of the design rules for the R-S codes are

$$q = 2^k$$

$$N = q - 1$$

$$2t = N - K$$

Here, t is the number of symbol errors that can be corrected. A simple example will be given to illustrate these. Let $k = 2$; then $q = 4$, and these four symbols may be labeled A, B, C, and D. In terms of the binary symbols (bits) for this simple case, we could have $A = 00$, $B = 01$, $C = 10$, and $D = 11$. One could imagine the binary numbers 00, 01, 10, and 11 being stored in memory locations labeled A, B, C, and D.

The number of symbols per codeword is $N = q - 1 = 3$. Suppose that $t = 1$; then the rule $2t = N - K$ gives $K = 1$; that is, there will be one symbol per dataword. Hence the number of datawords is $q^K = 4$ (i.e., A, B, C, or D), and the number of codewords is $q^N = 4^3 = 64$. These will include permissible words of the form AP_1P_2, BP_3P_4, CP_5P_6, and DP_7P_8,

where P_1P_2, and so on are the parity symbols selected by the encoding rules from the symbol alphabet A, B, C, and D. This process is illustrated in Fig. 11.1.

At the decoder, these are the only words that are recognized as being legitimate and can be decoded. The other possible codewords not formed by the rules but which may be formed by transmission errors will be detected as errors and corrected. It will be observed that a codeword consists of 6 bits, and one or more of these in error will result in a symbol error. The R-S code is capable of correcting this symbol error, which in this simple illustration means that a burst of up to 6 bit errors can be corrected.

R-S codes do not provide efficient error correction where the errors are randomly distributed as distinct from occurring in bursts (Taub and Schilling, 1986). To deal with this situation, codes may be joined together or concatenated, one providing for random error correction and one for burst error correction. Concatenated codes are described in Sec. 11.6. It should be noted that although the encoder and decoder in R-S codes operate at the symbol level, the signal may be transmitted as a bit stream, but it is also suitable for transmission with multilevel modulation, the levels being determined by the symbols. The code rate is $r_c = K/N$, and the code is denoted by (N, K). In practice, it is often the case that the symbols are bytes consisting of 8 bits; then $q = 2^8 = 256$, and $N = q - 1 = 255$. With $t = 8$, a NASA-standard (255, 239) R-S code results.

Shortened R-S codes employ values $N' = N - l$ and $K' = K - l$ and are denoted as (N', K'). For example, DirecTV (see Chap. 16) utilizes a shortened R-S code for which $l = 109$, and *digital video broadcast* (DVB) utilizes one for which $l = 51$ (Mead, 2000). These codes are designed to correct up to $t = 8$ symbol errors.

11.4 Convolution Codes

Convolution codes are also linear codes. A convolution encoder consists of a shift register which provides temporary storage and a shifting operation for the input bits and exclusive-OR logic circuits which generate the coded output from the bits currently held in the shift register.

In general, k data bits may be shifted into the register at once, and n code bits generated. In practice, it is often the case that $k = 1$ and $n = 2$, giving rise to a rate 1/2 code. A rate 1/2 encoder is illustrated in Fig. 11.2, and this will be used to explain the encoding operation.

Initially, the shift register holds all binary 0s. The input data bits are fed in continuously at a bit rate R_b, and the shift register is clocked at this rate. As the input moves through the register, the rightmost bit is shifted out so that there are always 3 bits held in the register. At the end of the message, three binary 0s are attached, to return the shift register

data bits	01	00	11	00	01	11	01	01	11	01	10	10
datawords	B	A	D	A	B	D	B	B	D	B	C	C
codewords	BP_3P_4	AP_1P_2	DP_7P_8	AP_1P_2	BP_3P_4	DP_7P_8	BP_3P_4	BP_3P_4	DP_7P_8	BP_3P_4	CP_5P_6	CP_5P_6

Figure 11.1 Symbol (Reed-Solomon) encoding.

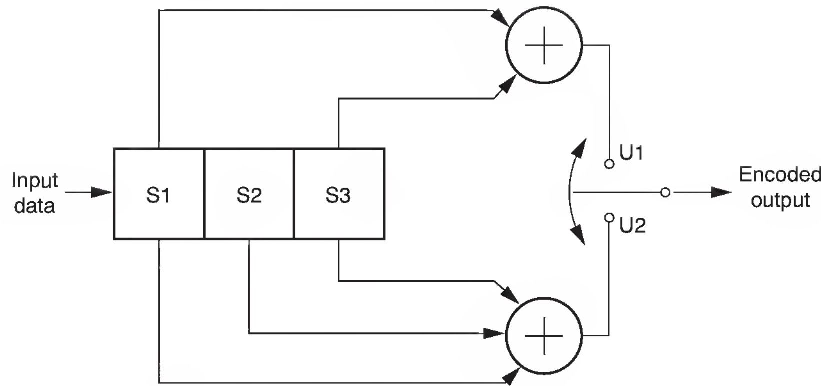


Figure 11.2 A rate 1/2 convolutional encoder.

to its initial condition. The commutator at the output is switched at twice the input bit rate so that two output bits are generated for each input bit shifted in. At any one time the register holds 3 bits which form the inputs to the exclusive-OR circuits.

Figure 11.3 is a tree diagram showing the changes in the shift register as input is moved in, with the corresponding output shown in parentheses. At the initial condition, the register stores 000, and the output is 00. If the first message bit in is a 1, the lower branch is followed, and the output is seen to be 11. Continuing with this example, suppose that the next three input bits are 001; then the corresponding output is 01 11 11. In other words, for an input 1001 (shown shaded in Fig. 11.3), the output, including the initial condition (enclosed here in brackets), is [00] 11 01 11 11. From this example it will be seen that any given input bit contributes, for as long as it remains in the shift register, to the encoded word. The number of stages in the register gives the constraint length of the encoder. Denoting the constraint length by m, the encoder is specified by (n, k, m). The example shows a (2, 1, 3) encoder. Encoders are optimized through computer simulation.

At the receiver, the tree diagram for the encoder is known. Decoding proceeds in the reverse manner. If, for example, [00] 11 01 11 11 is received, the tree is searched for the matching branches, from which the input can be deduced. Suppose, however, that an error occurs in transmission, changing the received word to [00] 01 01 11 11; i.e., the error is in the first bit following the initial condition. The receiver decoder expects either a 00 or a 11 to follow the initial 00; therefore, it has to make the assumption that an error has occurred. If it assumes that 00 was intended, it will follow the upper branch, but now a further difficulty arises. The next possible pair is 00 or 11, neither of which matches the received code word. On the other hand, if it assumes that 11 was

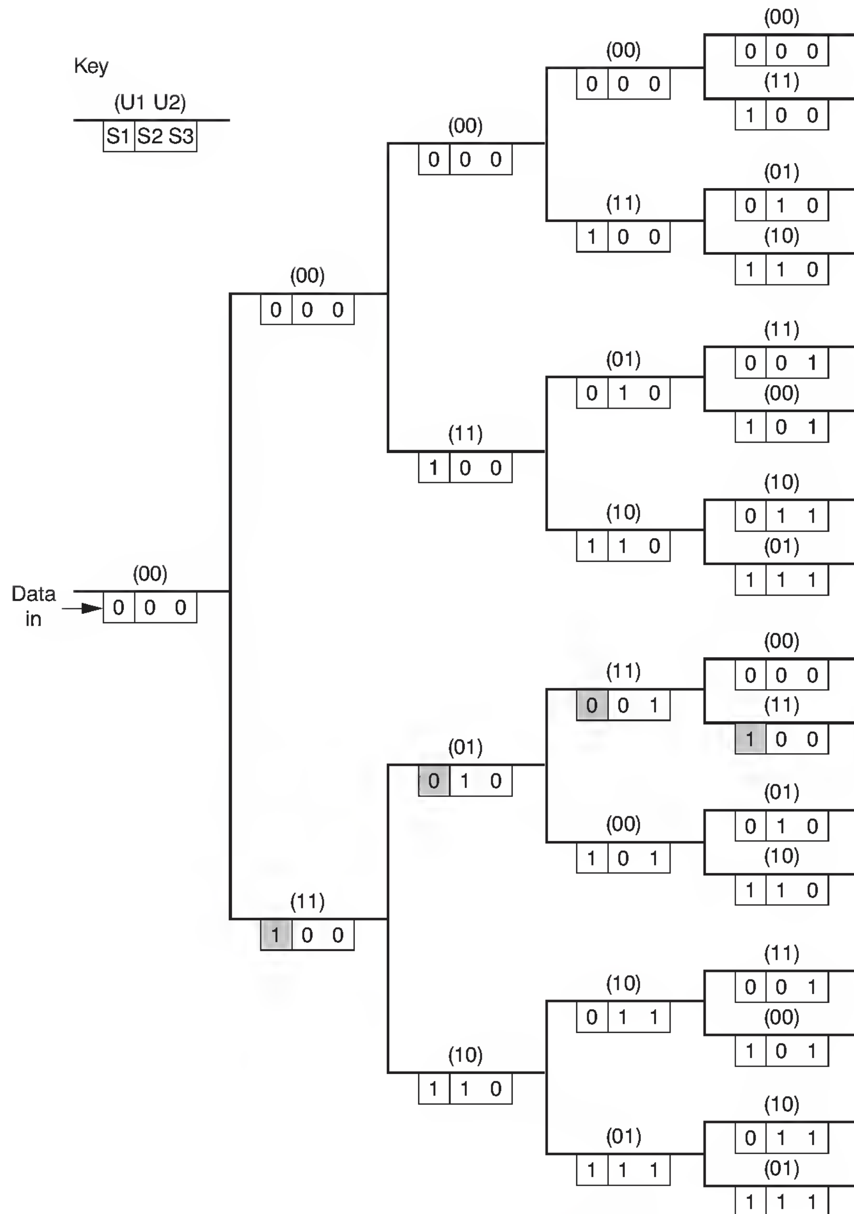


Figure 11.3 The tree diagram for the rate 1/2 convolutional encoder.

intended, it takes the lower branch, and then it can match all the following pairs with the branches in the decoding tree. On the basis of maximum likelihood, this would be the preferred path, and the correct input 1001 would be deduced.

The CDV-10MIC is a single integrated circuit which implements all the functions required for a constraint length 7, rate 1/2, and punctured 2/3 or 3/4 rate, convolutional encoder, and Viterbi algorithm decoder. Important features of this chip are:

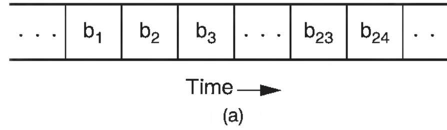
- Full decoder and encoder implementation for rates 1/2, 2/3, and 3/4
- Complies with INTELSAT IESS-308 and IESS-309 specifications
- Extremely low implementation margin
- No external components required for punctured code implementation
- Operates at all information rates up to 10 Mbits/s. Higher speed versions are under development
- All synchronization circuits are included on chip. External connection of ambiguity state counter and ambiguity resolution inputs allows maximum application flexibility
- Advanced synchronization detectors enable very rapid synchronization. Rate, 3/4 block and phase synchronization in less than 8200 information bits (5500 transmitted symbols).
- Soft decision decoder inputs (3 bits, 8 levels)
- Erasure inputs for implementing punctured codes at other rates
- Path memory length options to optimize performance when implementing high-rate punctured codes
- Error-monitoring facilities included on chip
- Synchronization detector outputs and control inputs to enable efficient synchronization in higher-speed multiplexed structures.

Figure 11.4 Specifications for a single-chip Viterbi codec. (*Courtesy of Signal Processors, Ltd., Cambridge U.K.*)

Decoding is a more difficult problem than encoding, and as the example suggests, the search process could quickly become impracticable for long messages. The Viterbi algorithm is used widely in practice for decoding. An example of a commercially available codec is the CDV-10MIC single-chip codec made by Signal Processors Limited, Cambridge, U.K. The data sheet for this codec is shown in Fig. 11.4. The CDV-10MIC utilizes Viterbi decoding. It has a constraint length of 7 and can be adjusted for code rates of 1/2, 2/3, and 3/4 by means of what is termed *punctured coding*. With punctured coding, the basic code is generated at code rate 1/2, but by selectively discarding some of the output bits, other rates can be achieved (Mead, 2000). The advantage is that a single encoder can be used for different rates.

11.5 Interleaving

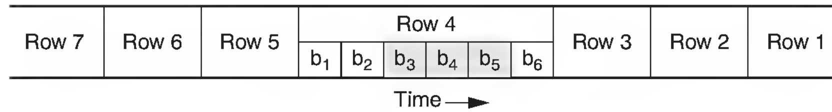
The idea behind interleaving is to change the order in which the bits are encoded so that a burst of errors gets dispersed across a number of codewords rather than being concentrated in one codeword. Interleaving as applied in block codes will be used here to illustrate



(a)

Column No. / Row No.	1	2	3	4	5	6
1	b_{24}	b_{23}	b_{22}	b_{21}	b_{20}	b_{19}
2	b_{18}	b_{17}	b_{16}	b_{15}	b_{14}	b_{13}
3	b_{12}	b_{11}	b_{10}	b_9	b_8	b_7
4	b_6	b_5	b_4	b_3	b_2	b_1
5	c_1	c_4	c_7	c_{10}	c_{13}	c_{16}
6	c_2	c_5	c_8	c_{11}	c_{14}	c_{17}
7	c_3	c_6	c_9	c_{12}	c_{15}	b_{18}

(b)



(c)

Figure 11.5 Illustrating interleaving (see Sec. 11.5).

the technique, but it also can be used with convolutional coding (Taub and Schilling, 1986).

Figure 11.5*a* shows part of the data bit stream where for definiteness the bits are labeled from b_1 to b_{24}. These are fed into shift registers as shown in Fig. 11.5*b*, where, again, for definiteness seven rows and six columns are shown. Rather than encoding the rows, the columns are encoded so that the parity bits fill up the last three rows. It will be seen, therefore, that the bits are not encoded in the order in which they appear in the data bit stream. The encoded bits are read out row by row as shown in Fig. 11.5*c*. Row 4 is shown in detail. If now an error burst occurs which changes bits b_5, b_4, and b_3, these will appear as separate errors in the encoded words formed by columns 2, 3, and 4. The words formed by the column bits are encoded to correct single errors (in this example), and therefore, the burst of errors has been corrected.

11.6 Concatenated Codes

Codes designed to correct for burst errors can be combined with codes designed to correct for random errors, a process known as *concatenation*. Figure 11.6 shows the general form for concatenated codes. The input data are fed into the encoder designed for burst error correction. This is the outer encoder. The output from the outer encoder is fed into the encoder designed for random error correction. This is the inner encoder. The signal is then modulated and passed on for transmission. At the receiver, the signal is demodulated. The inner decoder matches the inner encoder and follows the demodulator. The output from the inner decoder is fed into the outer decoder, which matches the outer encoder. The term *outer* refers to the outermost encoding/decoding units in the equipment chain, and the term *inner* refers to the innermost encoding/decoding unit. In digital satellite television, the outer code is a R-S code, and the inner code is a convolutional code. The inner decoder utilizes Viterbi decoding.

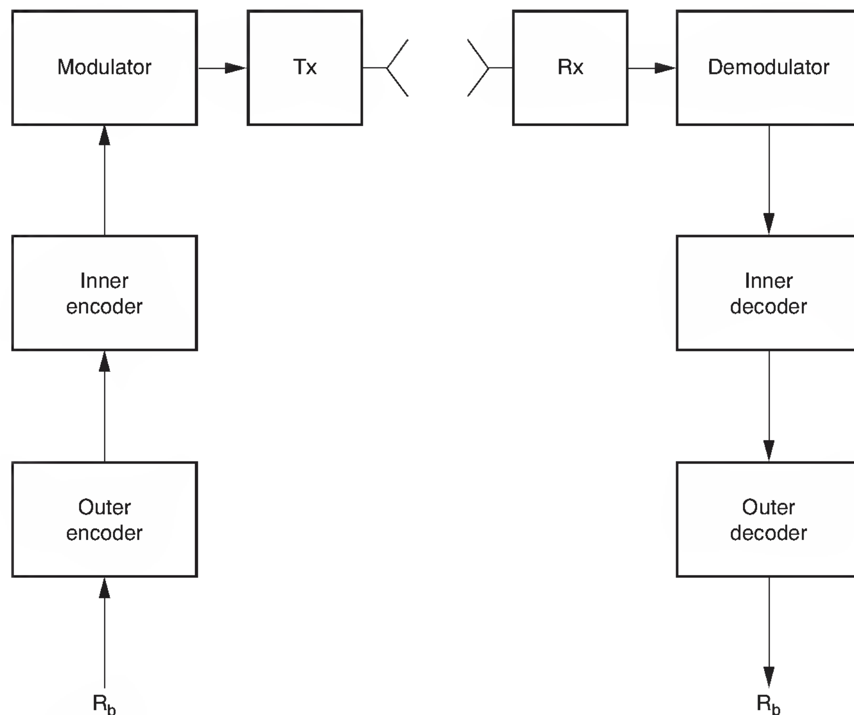


Figure 11.6 Concatenated coding (see Sec. 11.6).

11.7 Link Parameters Affected by Coding

Where no error control coding is employed, the message will be referred to as an *uncoded message*, and its parameters will be denoted by the subscript U. Figure 11.7*a* shows the arrangement for an uncoded message. Where error control coding is employed, the message will be referred to as a *coded message*, and its parameters will be denoted by the subscript C. Figure 11.7*b* shows the arrangement for a coded message. For comparison purposes, the $[C/N_0]$ value is assumed to be the same for both situations. The input bit rate to the modulator for the uncoded message is R_b, and for the coded message is R_c. Since n code bits must be transmitted for every k data bits, the bit rates are related as

$$\frac{R_b}{R_c} = r_c \tag{11.10}$$

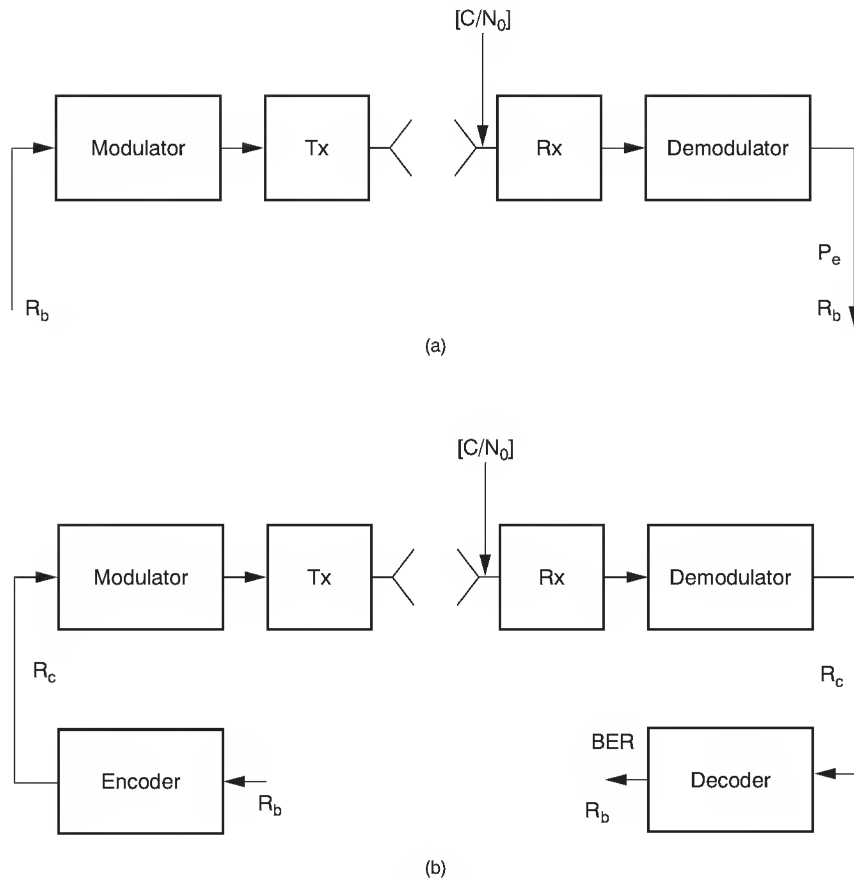


Figure 11.7 Comparing links with and without FEC.

Since r_c is always less than unity, then $R_c > R_b$ always. For constant carrier power, the bit energy is inversely proportional to bit rate (see Eq. 10.22), and therefore,

$$\frac{E_c}{E_b} = r_c \tag{11.11}$$

where E_b is the average bit energy in the uncoded bit stream (as introduced in Chap. 10), and E_c is the average bit energy in the coded bit stream.

Equation (10.18) gives the probability of bit error for *binary phase-shift keying* (BPSK) and *quadrature phase-shift keying* (QPSK) modulation. With no coding applied, E_b is just the E_b of Eq. (10.18), and the probability of bit error in the uncoded bit stream is

$$P_{eU} = \frac{1}{2}\operatorname{erfc}\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{11.12}$$

For the coded bit stream, the bit energy is $E_c = r_c E_b$, and therefore, Eq. (10.18) becomes

$$P_{eC} = \frac{1}{2}\operatorname{erfc}\left(\sqrt{\frac{r_c E_b}{N_0}}\right) \tag{11.13}$$

This means that $P_{eC} > P_{eU}$, or the probability of bit error with coding is worse than that without coding. It is important to note, however, that the probability of bit error applies at the input to the decoder. For the error control coding to be effective, the output BER should be better than that obtained without coding. More will be discussed about this later.

The limitation imposed by bandwidth also must be considered. If the time for transmission is to be the same for the coded message as for the uncoded message, the bandwidth has to be increased to accommodate the higher bit rate. The required bandwidth is directly proportional to bit rate (see Eq. 10.16), and hence it has to be increased by a factor $1/r_c$. If, however, the bandwidth is fixed (the system is band limited), then the only recourse is to increase the transmission time by the factor $1/r_c$. For a fixed number of bits in the original message, the bit rate R_b entering into the encoder is reduced by a factor r_c compared with what it could have been without coding.

As an example, it is shown in Sec. 10.4 that the TI message rate is 1.544 Mb/s. When 7/8 FEC is applied, the transmission rate becomes

$1.544 \times 8/7 = 1.765$ Mb/s. From Eq. (10.16), the required bandwidth is $B_{IF} = 1.765 \times (1.2)/2 = 1.06$ MHz.

11.8 Coding Gain

As shown by Eqs. (11.12) and (11.13), the probability of bit error for a coded message is higher (therefore, worse) than that for an uncoded message, and therefore, to be of advantage, the coding itself must more than offset this reduction in performance. In order to illustrate this, the messages will be assumed to be BPSK (or QPSK) so that the expressions for error probabilities as given by Eqs. (11.12) and (11.13) can be used. Denoting by BER_U the bit error rate after demodulation for the uncoded message and by BER_C the bit error rate for the coded message after demodulation and decoding, then for the uncoded message

$$\mathrm{BER}_U = P_{eU} \tag{11.14}$$

Certain codes known as *perfect codes* can correct errors up to some number t. The BER for such codes is given by (see Roddy and Coolen, 1995)

$$\mathrm{BER}_C = \frac{(n-1)!}{t!(n-1-t)!} P_{eC}^{t+1} \tag{11.15}$$

where $x! = x(x-1)(x-2)\ldots 3.2.1$ (and n is the number of bits in a codeword). The Hamming codes are perfect codes that can correct one error. For this class of codes and with $t = 1$, Eq. (11.15) simplifies to

$$\mathrm{BER}_C = (n-1)P_{eC}^2 \tag{11.16}$$

A plot of BER_C and BER_U against $[E_b/N_0]$ is shown in Fig. 11.8 for the Hamming (7, 4) code. The crossover point occurs at about 4 dB, so for the coding to be effective, $[E_b/N_0]$ must be higher than this. Also, from the graph, for a BER of 10^{-5}, the $[E_b/N_0]$ is 9.6 dB for the uncoded message and 9 dB for the coded message. Therefore, at this BER value the Hamming code is said to provide a coding gain of 0.6 dB.

Some values for coding gains given in Taub and Schilling (1986) are block codes, 3 to 5 dB; convolutional coding with Viterbi decoding, 4 to 5.5 dB; concatenated codes using R-S block codes and convolutional decoding with Viterbi decoding, 6.5 to 7.5 dB. These values are for a P_e value of 10^{-5} and using hard decision decoding as described in the following section.

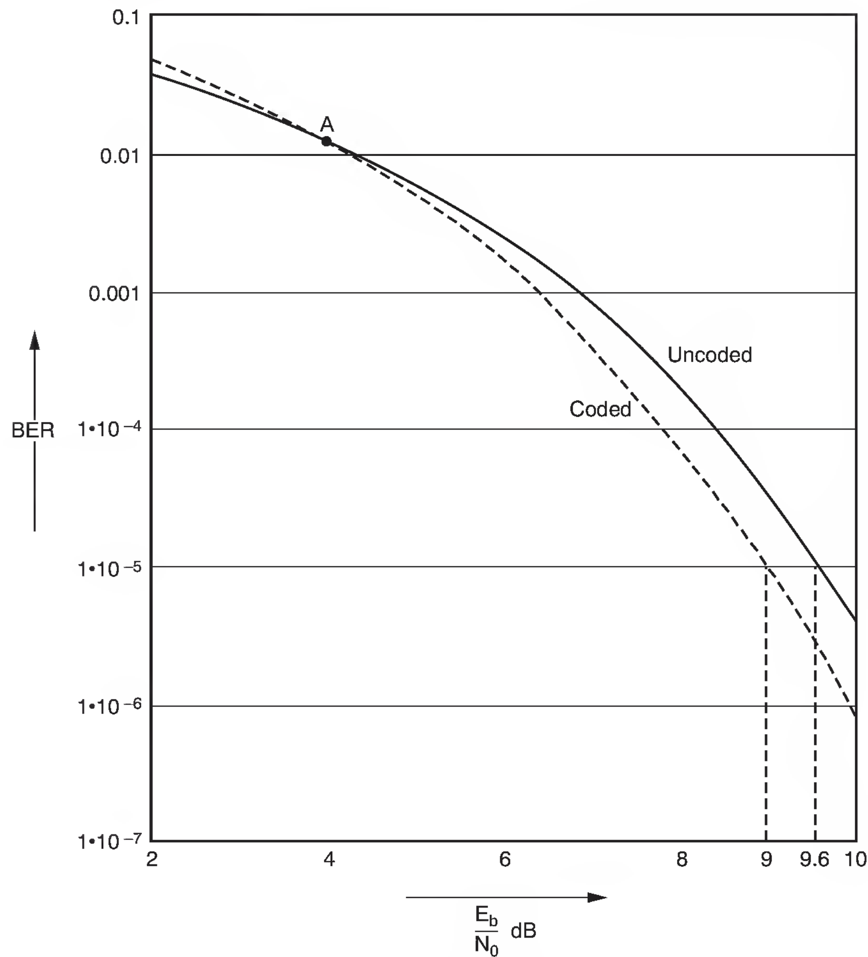


Figure 11.8 Plot of BER versus $[E_b/N_0]$ for coded and uncoded signals.

11.9 Hard Decision and Soft Decision Decoding

With hard decision decoding, the output from the optimum demodulator is passed to a threshold detector that generates a "clean" signal, as shown in Fig. 11.9*a*. Using triple redundancy again as an example, the two codewords would be 111 and 000. For binary polar signals, these might be represented by voltage levels 1 V 1 V 1 V and −1 V −1 V −1 V. The threshold level for the threshold detector would be set at 0 V. If now the sampled signal from the optimum demodulator is 0.5 V 0.7 V −2 V, the output from the threshold detector would be 1 V 1 V −1 V, and the decoder would decide that this was a binary 1 1 0 codeword and produce a binary 1 as

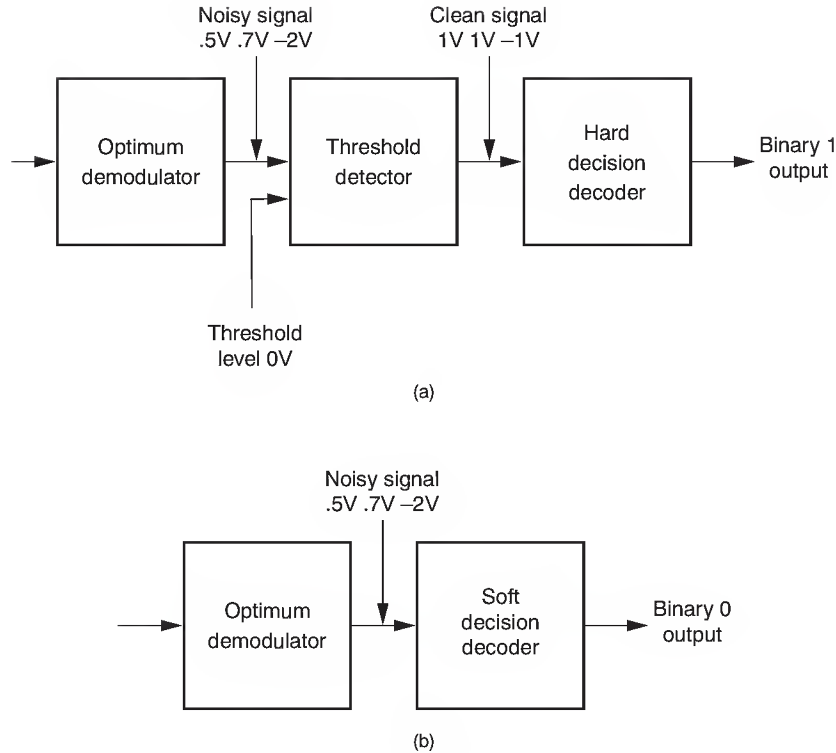


Figure 11.9 (*a*) Hard decision and (*b*) soft decision decoding.

output. In other words, a firm or hard decision is made on each bit at the threshold detector.

With soft decision decoding (Fig. 11.9*b*), the received codeword is compared in total with the known codewords in the set, 111 and 000 in this example. The comparison is made on the basis of minimum distance (the minimum distance referred to here is a *Euclidean distance* as described shortly. This is not the same as the minimum distance introduced in Sec. 11.2). To illustrate this, consider the first two points in an x, y, z coordinate system. Let point P_1 have coordinates x_1, y_1, z_1 and point P_2 have coordinates x_2, y_2, z_2. From the geometry of the situation, the distance d between the points is obtained from

$$d^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2$$

Treating the codewords as vectors and comparing the received codeword on this basis with the stored version of 111 results in

$$(0.5 - 1)^2 + (0.7 - 1)^2 + (-2 - 1)^2 = 9.34$$

Comparing it with the stored version of 000 results in

$$[0.5 - (-1)]^2 + [0.7 - (-1)]^2 + [-2 - (-1)]^2 = 6.14$$

The distance determined in this manner is often referred to as the *Euclidean distance* in acknowledgment of its geometric origins, and the distance squared is known as the *Euclidean distance metric*. On this basis, the received codeword is closest to the 000 codeword, and the decoder would produce a binary 0 output.

Soft decision decoding results in about a 2-dB reduction in the $[E_b/N_0]$ required for a given BER (Taub and Schilling, 1986). This reference also gives a table of comparative values for soft and hard decision coding for various block and convolutional codes. Clearly, soft decision decoding is more complex to implement than hard decision decoding and is only used where the improvement it provides must be had.

11.10 Shannon Capacity

In a paper on the mathematical theory of communication (Shannon, 1948) Shannon showed that the probability of bit error could be made arbitrarily small by limiting the bit rate R_b to less than (and at most equal to) the *channel capacity*, denoted by C. Thus

$$R_b \leq C \tag{11.17}$$

For random noise where the spectrum density is flat (this is the N_0 spectral density previously introduced) the channel capacity is given by

$$C = W \log_2\left(1 + \frac{S}{N}\right) \tag{11.18}$$

Here, W is the baseband bandwidth, and S/N is the baseband signal to noise power ratio (not decibels). Shannon's theorem can be written as

$$R_b \leq W \log_2\left(1 + \frac{S}{N}\right) \tag{11.19}$$

Letting P_R represent the average signal power, and T_b the bit period then as shown by Eq. (10.17) the bit energy is $E_b = P_R T_b$. The noise power is $N = WN_0$ and the signal to noise ratio is

$$\begin{aligned} \frac{S}{N} &= \frac{P_R}{N} \\ &= \frac{E_b}{T_b W N_0} \end{aligned} \tag{11.20}$$

The bit rate is $R_b = 1/T_b$. Substituting this in Eq. (11.20) then in the inequality (11.19) gives

$$\frac{R_b}{W} \leq \log_2\left(1 + \frac{R_b}{W}\frac{E_b}{N_0}\right) \tag{11.21}$$

As noted in Sec. 10.6.3, the ratio of bit rate to bandwidth (R_b/W in the inequality (11.21) is an important characteristic of any digital system. The greater this ratio, the more efficient the system. The limiting case is when the inequality sign is replaced by the equal sign.

$$\frac{R_b}{W} = \log_2\left(1 + \frac{R_b}{W}\frac{E_b}{N_0}\right) \tag{11.22}$$

Keep in mind that this relationship is for the condition of arbitrarily small probability of bit error. A plot of R_b/W as a function of E_b/N_0 is shown in

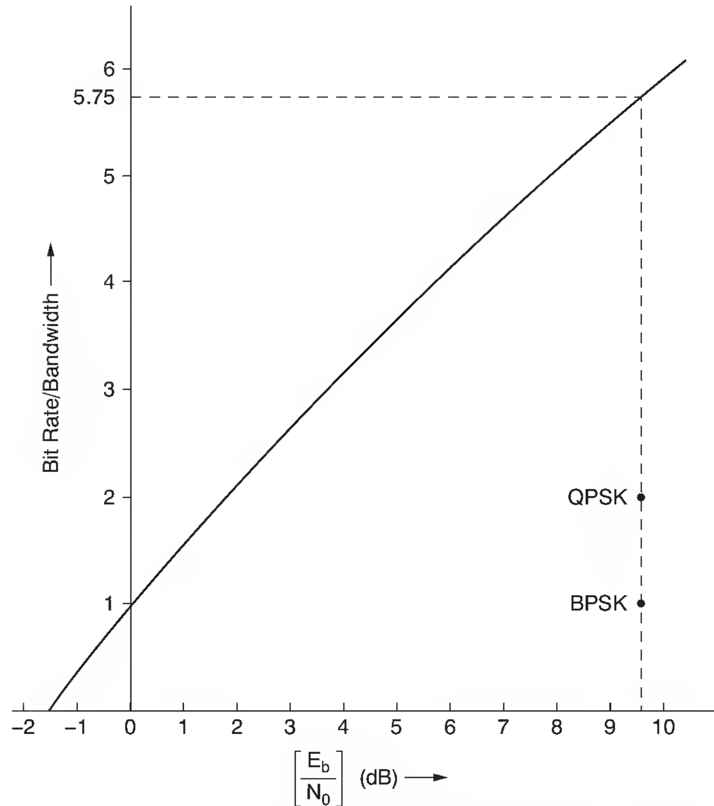


Figure 11.10 Graph showing the Shannon limit, Eq. (11.22). The units for the y-axis are (bits/s)/Hz (a dimensionless ratio in fact). The points for BPSK and QPSK are evaluated for a $P_e = 10^{-5}$.

Fig. 11.10. Note that although the graph shows E_b/N_0 in decibels ($[E_b/N_0]$ in our previous notation) the power ratio must be used in evaluating Eq. (11.22).

In any practical system there will be a finite probability of bit error, and to see how this fits in with the Shannon limit, consider the BER graph of Fig. 10.17, which applies for BPSK and QPSK. From Fig. 10.17 [or from calculation using Eq. (10.18)], for a probability of bit error (or BER in this case) of 10^{-5} the $[E_b/N_0]$ is about 9.6 dB. (See also the uncoded curve of Fig. 11.8). As shown in Sec. 10.6.3 the bit rate to bandwidth ratio is $1/(1 + \rho)$ for BPSK and $2/(1 + \rho)$ for QPSK. For purposes of comparison ideal filtering will be assumed, for which $\rho = 0$. Thus on Fig. 11.10, the points (1, [9.6]) for BPSK and (2, [9.6]) can be shown. At an $[E_b/N_0]$ of 9.6 dB the Shannon limit indicates that a bit rate to bandwidth ratio of about 5.75 : 1 should be achievable, and it is seen that BPSK and QPSK are well below this. Alternatively, for a bit-rate/bandwidth ratio of 1, the Shannon limit is 0 dB (or an E_b/N_0 ratio of unity), compared to 9.6 dB for BPSK.

11.11 Turbo Codes and LDPC Codes

Till 1993 all codes used in practice fell well below the Shannon limit. In 1993, a paper (Berrou et al., 1993) presented at the IEEE International Conference on Communications made the claim for a digital coding method that closely approached the Shannon limit (a pdf file for the paper will be found at www-elec.enst-bretagne.fr/equipe/berrou/ Near%20Shannon%20Limit%20Error.pdf). Subsequent testing confirmed the claim to be true. This revitalized research into coding, resulting in a number of "turbo-like" codes, and a renewed interest in codes known as *low density parity check* (LDPC) codes (see Summers, 2004).

Turbo codes and LDPC codes use the principle of *iterative decoding* in which "soft decisions" (i.e., a probabilistic measure of the binary 1 or 0 level) obtained from different encoding streams for the same data, are compared and reassessed, the process being repeated a number of times (iterative processing). This is sometimes referred to as *soft input soft output* to describe the fact that during the iterative process no hard decisions (binary 1 or 0) are made regarding a bit. Each reassessment generally provides a better estimate of the actual bit level, and after a certain number of iterations (fixed either by convergence to a final value, or by a time limit placed on the process) a hard decision output is generated.

Turbo codes are so named because the iterative or feedback process was likened to the feedback process in a turbo-charged engine (see Berrou et al., 1993). The turbo principle can be applied with concatenated block

codes, (known as *Turbo Product codes* or TPCs, see Comtech, 2002). However the more common arrangement is to use parallel concatenation using convolution encoders. Because of the continuous nature of convolution coding, data and code *sequences* rather than words are involved. The convolution encoders shown in Fig. 11.11 (Burr, 2001) are *recursive convolution encoders*. They differ from the convolution encoder of Fig. 11.2 in that feedback is employed. (This *recursive* feedback is part of the encoding process and is quite separate from the iterative feedback to be described shortly, which gives turbo codes their name). It can be shown that the recursive feedback (see Burr, 2001) assists in maintaining a large minimum Hamming distance (see Sec. 11.2) for code sequences. Another difference between the encoders of Fig. 11.11 and that of Fig. 11.2 is that the data sequence is fed directly to the multiplexer of Fig. 11.11, making the output systematic (see Sec. 11.2).

Parity-1 bits are generated directly from the data bits and parity-2 bits from the interleaved data bits, so that two independent streams of parity bits are generated. Interleaving as described in Sec. 11.5, was used there as a means of combating bursty errors. With turbo encoding the purpose of interleaving is different, it is used to provide independent parity bits for the same input. A number of different methods of interleaving are available, and the design of the interleaver is a crucial aspect of turbo code design.

The output coded sequence is *data, parity-1, parity-2*. Since each encoder generates a parity bit for every data bit the code rate is 1/3. This is relatively low but can be increased by puncturing, as described in Sec. 11.4 and shown in Fig. 11.11. For example, one parity bit might be discarded in turn from each of the encoders, resulting in a 1/2 code rate. Other rates are possible with puncturing. If puncturing is used with the encoder, dummy bits are inserted at the decoder to replace

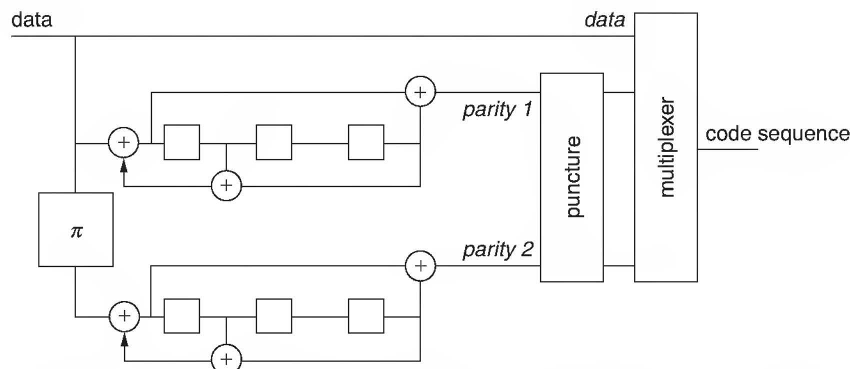


Figure 11.11 A turbo encoder with puncturing. The π symbol indicates an interleaver, and the $\oplus$ symbol indicates modulo-2 addition. (*Courtesy of A Burr and the IEEE.*)

those discarded. The dummy bit level is set midway between the binary 1 and 0 levels, so they do not affect the decoding process.

The block schematic for the decoder is shown in Fig. 11.12 (Burr, 2001). The demultiplexer provides outputs for the data sequence and parity-1 and parity-2 sequences. These outputs are "soft," meaning that some measure of the bit level is used rather than a hard decision output of binary 1 or 0. For example, assuming a threshold decision level of 0.5V, the demultiplexer output might be 0.9V, 0.7V 0.1V, 0.2V, 0.9V, 0.65V, 0.3V, suggesting a hard decision binary sequence of 1 1 0 0 1 1 0. However the hard decision output does not make use of the *likelihood* of the hard decision being correct. The 0.9V level is obviously more likely to be a binary1 than the 0.65V level. A statistical measure termed the *log-likelihood ratio* (LLR) is most commonly used. For a given received value r, let $p(1/r)$ represent the probability that a 1 was transmitted, and $p(0/r)$ the probability that a 0 was transmitted. The log likelihood ratio is defined as

$$\text{LLR} = \log_e\left(\frac{p(1/r)}{p(0/r)}\right) \tag{11.23}$$

Where the transmission of 1s and 0s are equiprobable, (the probability of either a 1 or 0 occurring being 1/2 , rather like the probability of heads or tails of a the toss of a fair coin) the LLR becomes:

$$\begin{aligned}\text{LLR} &= \log_e\left(\frac{p(r/1)}{p(r/0)}\right) \\ &= \log_e p(r/1) - \log_e p(r/0)\end{aligned} \tag{11.24}$$

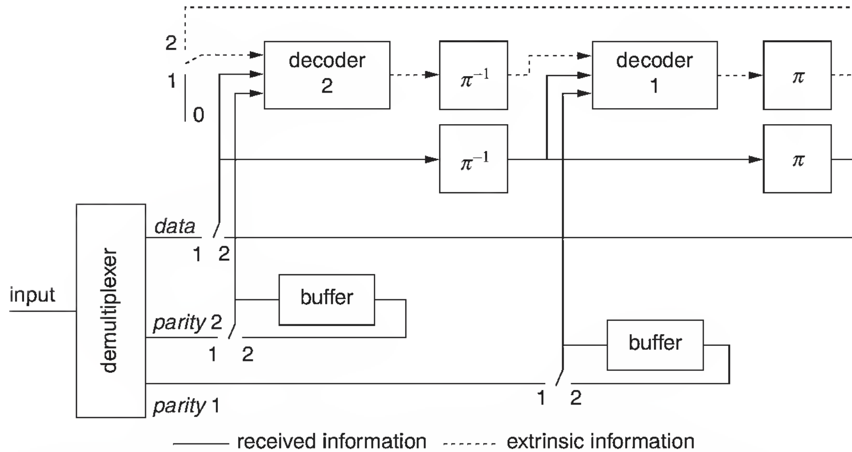


Figure 11.12 The turbo decoder for the encoder of Figure 11.11. The symbol π^{-1} represents a deinterleaver. (*Courtesy of A Burr and the IEEE.*)

where $p(r/1)$ is the probability of receiving value r, given that a 1 was transmitted and $p(r/0)$ the probability of receiving value r, given that a 0 was transmitted. If the voltage levels are normalized so that 1V represents a probability of 1, a certainty and 0V, zero probability, then for $r = 0.9$V for example, $p(r/1) = 0.9$ and $p(r/0) = 1 - .9 = 0.1$, so that LLR $= 2.197$. With $r = 0.3$, LLR $= -0.847$. In general, LLR yields a positive number for r closer to 1 and a negative number for r closer to 0. The magnitude of LLR is a measure of "how close." These two pieces of information are included in the soft sequences that form a part of the output of the multiplexer and which is the input to the decoders. The outputs from the decoders are also "soft" and the system is referred to as *soft-input soft-output* (SISO).

As shown in Fig. 11.12, the switches are in position 1 for the first iteration of the decoding step. Following the first iteration the switches are switched to position 2, and each decoder makes use of the soft information obtained from the other decoder to obtain a better estimate of bit values. Recall that two independent parity sequences are available for a given data sequence. The decoded data is adjusted to take into account the new estimates, and the process is repeated a number of times, typically for 4 to 10, before a final hard decision is made. The information that is obtained from the received data bits is termed *intrinsic information*, the intrinsic information flow paths being shown by the solid line in Fig. 11.12. The information that is passed from one decoder to the other is termed *extrinsic information*, the paths for the extrinsic information flow being shown by the dotted line. After the final iteration the output of the second decoder is switched to the output line (not shown in Fig. 11.12). It will be a 1 for a positive LLR and a 0 for a negative LLR. A more detailed description of the encoder of Fig. 11.11 and the decoder of Fig. 11.12 will be found in Burr (2001).

11.11.1 Low density parity check (LDPC) codes

LDPC refers to the fact that the parity check matrix (Sec. 11.2) is sparse, that is, it has few binary 1s compared to binary 0s. The LDPC codes were first introduced by Gallagher (1962) who showed that a low density parity check matrix resulted in excellent minimum distance properties (as defined in Sec. 11.2), and they are comparatively easy to implement. As mentioned above a feature common to LDPC codes and turbo codes is that SISO decoding is employed, and a series of iterations performed to (hopefully) improve the probability estimate of a bit being a 1 or 0. Only after a predetermined number of iterations is a "hard decision" arrived at.

An example of a parity check matrix for a LDPC code (Summers, 2004) is

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{11.25}$$

As shown in connection with Eq. (11.5) the number of rows in **H** is equal to the number of parity bits $n-k$, and the number of columns is equal to the length n of the codeword. In this case $n-k = 7$ and $n = 16$, hence $k = 9$, and the **H** matrix represents a (16, 9) code. From Eq. (11.7) the syndrome is obtained on multiplying the received codeword by $\mathbf{H}^{\mathbf{T}}$, the transpose of **H**, and ideally, an error-free codeword is indicated by an all-zero syndrome. Standard practice is to index bit positions starting from zero, thus a 16-bit codeword would have the bits labeled c_0, c_1, $c_2, \ldots c_{15}$. Likewise, elements in the **H** matrix are labeled h_{pq} where the first element (top left-hand corner) is h_{00}.

In general the row number (indexed from zero) gives the number of the syndrome element, and the 1s in the columns indicate which codeword bits are used. The seven parity check equations obtained from the **H** matrix, are, on setting the syndrome equal to 0.

$$\begin{aligned} c_0 \oplus c_1 \oplus c_2 \oplus c_9 &= 0 \\ c_3 \oplus c_4 \oplus c_5 \oplus c_{10} &= 0 \\ c_6 \oplus c_7 \oplus c_8 \oplus c_{11} &= 0 \\ c_0 \oplus c_3 \oplus c_6 \oplus c_{12} &= 0 \\ c_1 \oplus c_4 \oplus c_7 \oplus c_{13} &= 0 \\ c_2 \oplus c_5 \oplus c_8 \oplus c_{14} &= 0 \\ c_{12} \oplus c_{13} \oplus c_{14} \oplus c_{15} &= 0 \end{aligned} \tag{11.26}$$

As noted in connection with Eq. (11.3) a systematic code has the dataword at the beginning of the codeword, thus it follows that the columns 0 to 8 of the **H** matrix operate on the datawords. The fact that each column has two 1s means that two of the dataword bits appear in each parity check equation determined by these columns. A standard way of showing the parity check equations and the codeword bits is by means of a *Tanner graph* (Tanner, 1981) in which circles represent the bit nodes and squares represent the parity check equations, Fig. 11.13. The

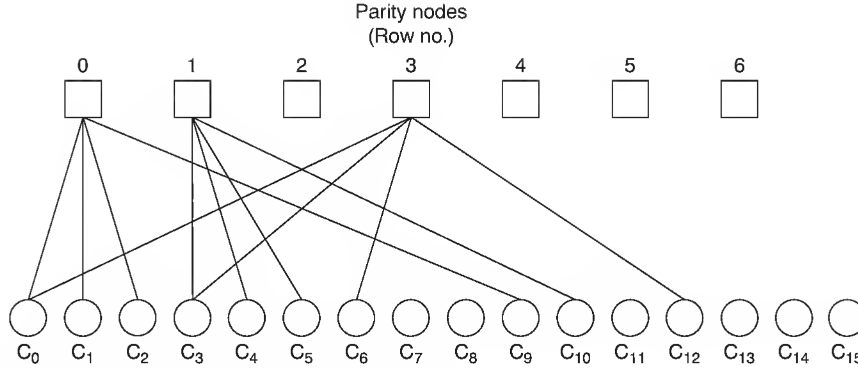


Figure 11.13 Illustrating a Tanner graph.

lines (technically referred to as *edges*) join the bit nodes to their respective parity check nodes. Edges occur wherever a 1 appears in the **H** matrix. Thus for row 0, 1s appear in the c_0, c_1, c_2 and c_9 positions [(and as shown by the first parity line in Eq. (11.26)]. In Fig. 11.13, only the parity equations for rows 0, 1, and 3 are shown for clarity, but the complete Tanner graph would show all the edges. Messages pass along the edges. Initially, the output from the channel demodulator provides the first "soft" estimate of a bit. If p^1 is the probability that, the bit is a 1, the probability that it is a zero is $1 - p^1$. These estimates are sent to their respective parity check equations where the equation is used to derive probability estimates for a bit. Considering the first equation for example, an estimate for the probability that c_0 is a 1 can be obtained from the probabilities for c_1, c_2, and c_9 being 1. For the group $c_1c_2c_9$ the combinations that would result in c_0 being 1 are 100, 010, 100, and 111. The sum of the corresponding probabilities gives an estimate, p'_0 for the probability of c_0 being 1:

$$p'_0 = p_1^1\left(1 - p_2^1\right)\left(1 - p_9^1\right) + p_2^1\left(1 - p_1^1\right)\left(1 - p_9^1\right) + p_9^1\left(1 - p_1^1\right)\left(1 - p_2^1\right) + p_1^1 p_2^1 p_9^1 \tag{11.27}$$

The estimates from the parity nodes are returned to the respective bit nodes. The bit node now has estimates from the parity check nodes and from the channel, which enables a new estimate for probability to be calculated. For example if bit node 0 receives estimates from parity check nodes A, B, and C, denoted by p_A^1, p_B^1, p_C^1 respectively, and p_{CH}^1 from the channel, the new estimates sent to these parity check nodes are the products $K(p_{CH}^1 p_B^1 p_C^1)$ to parity node A, $K(p_{CH}^1 p_A^1 p_C^1)$ to parity node B, and $K(p_{CH}^1 p_A p_B)$ to parity node C, where K is a normalizing constant. It will

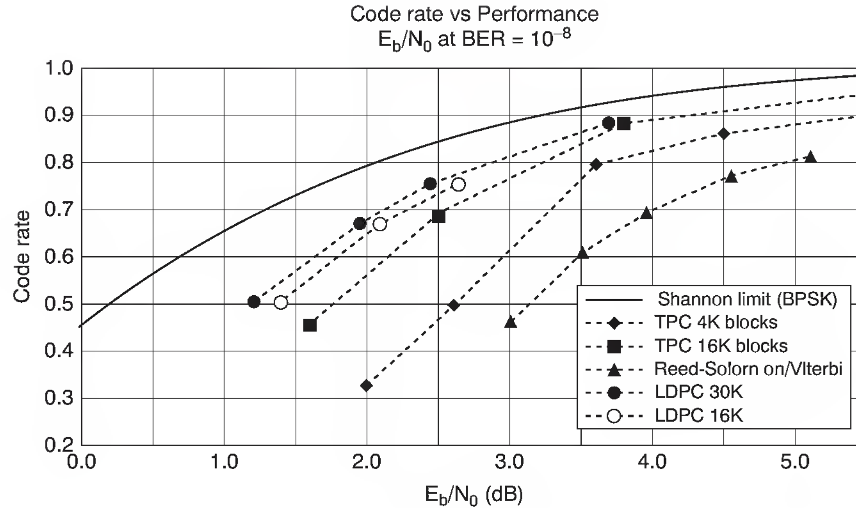


Figure 11.14 FEC performance of Comtech AHA 4701 LDPC coder. (*Courtesy of A. Summers Comtech AHA.*)

be noticed that the estimate from any given parity node is not included in the new estimate sent to that parity node. This process is repeated a number of times (iterated) before a final "hard decision" is made.

Turbo codes and LDPC codes are being employed in a number of satellite systems with a resulting improvement in channel performance, for example the *Digital Video Broadcast S2 standard* (DVB-S2) employs LDPC as the inner code in its FEC arrangement (Breynaert, 2005, Yoshida, 2003), and the DVB-RCS plans to use turbo codes (talk Satellite, 2004). The performance of a number of codes is shown in Fig. 11.14.

11.12 Automatic Repeat Request (ARQ)

Error detection without correction is more efficient than FEC in terms of code utilization. It is less complicated to implement, and more errors can be detected than corrected. Of course, it then becomes necessary for the receiver to request a retransmission when an error has been detected. This is an automatic procedure, termed *automatic repeat request* (ARQ). The request for retransmission must be made over a low-bit-rate channel where the probability of bit error can be kept negligibly small. Because of the long round-trip delay time, on the order of half a second or more, encountered with geostationary satellites, ARQ is only suitable for transmission that is not sensitive to long delays. ARQ is normally used with block encoding.

An estimate of the probability of an error remaining undetected can be made. With an (n, k) block code, the number of datawords is 2^k, and

the total number of codewords is 2^n. When a given dataword is transmitted, an undetected error results when the transmission errors convert the received codeword into one that contains a permissible dataword but not the one that was transmitted. The number of such datawords is $2^k - 1$. An upper bound on the probability of an error getting through can be made by assuming that all codewords are equiprobable. The ratio of number of possible error words to total number of codewords then gives the average probability of error. In practice, of course, all the codewords will not be equiprobable, those containing datawords being more probable than those which do not. The ratio, therefore, gives an upper bound on the probability of error:

$$\begin{aligned} BER &\leq \frac{2^k - 1}{2^n} \\ &< 2^{-(n-k)} \end{aligned} \tag{11.28}$$

where $n - k$ is the number of redundant bits in a codeword. For example, a (15, 11) code has an upper bound of approximately 0.06, while a (64, 32) code has an upper bound of approximately 2.3×10^{-10}.

It will be assumed, therefore, that the data are sent in coded blocks (referred to simply as *blocks* in the following). The receiver acknowledges receipt of each block by sending back a positive acknowledgment or ACK signal if no errors are detected in the block and a negative acknowledgment or NAK signal if errors are detected. In what is termed *stop and wait* ARQ, the transmitter stores a copy of the block just transmitted and waits for the acknowledgment signal. If a NAK signal is received, it retransmits the block, and if an ACK signals is received, it transmits the next block. In either case, the delay between transmissions is about half a second, the round-trip time to and from a geostationary satellite. This would be unacceptable in many applications.

What is required is continuous transmission of blocks incorporating the retransmission of corrupted blocks when these are detected. *Go back N ARQ* achieves this by having the blocks and the acknowledgment signals numbered. The transmitter must now be capable of storing the number of blocks N transmitted over the round-trip time and updating the storage as each ACK signal is received. If the receiver detects an error in block i, say, it transmits an NAK_i signal and refuses to accept any further blocks until it has received the correct version of block i. The transmitter goes back to block i and restarts the transmission from there. This means that block i and all subsequent blocks are retransmitted. It is clear that a delay will only be encountered when an NAK signal is received, but there is the additional time loss resulting from the retransmission of the good blocks following the corrupted block. The method can be further improved by using what is termed *selective*

repeat ARQ, where only a correct version of the corrupted block is retransmitted, and not the subsequent blocks. This creates a storage requirement at the receiver because it must be able to store the subsequent blocks while waiting for the corrected version of the corrupted block.

It should be noted that in addition to the ACK and NAK signals, the transmitter operates a timeout clock. If the acknowledgement signal (ACK or NAK) for a given block is not received within the timeout period, the transmitter puts the ARQ mechanism into operation. This requires the receiver to be able to identify the blocks so that it recognizes which ones are repeats.

The *throughput* of an ARQ system is defined at the ratio of the average number of data bits accepted at the receiver in a given time to the number of data bits that could have been accepted if ARQ had not been used. Let P_A be the probability that a block is accepted; then, as shown in Taub and Schilling (1986), the throughputs are

$$\text{Go back } N = \frac{k}{n}\frac{P_A}{P_A + N(1 - P_A)} \tag{11.29}$$

$$\text{Selective repeat } = \frac{k}{n}P_A \tag{11.30}$$

Typically, for a BCH (1023, 973) code and $N = 4$, the throughput for go back N is 0.915, and for selective repeat, 0.942.

It is also possible to combine ARQ methods with FEC in what are termed *hybrid ARQ systems*. Some details will be found in Pratt and Bostian (1986).

11.13 Problems and Exercises

11.1. Explain in your own words how *error detection* and *error correction* differ. Why would FEC normally be used on satellite circuits?

11.2. A transmission takes place where the average probability of bit error is 10^{-6}. Given that a message containing 10^8 bits is transmitted, what is the average number of bit errors to be expected?

11.3. A transmission utilizes a code for which $k = 7$ and $n = 8$. How many codewords and how many datawords are possible?

11.4. A triple redundancy coding scheme is used on a transmission where the probability of bit error is 10^{-5}. Calculate the probability of a bit error occurring in the output when (*a*) error detection only is used and (*b*) when error correction with majority voting is implemented.

11.5. Using the generator matrix given in Eq. (11.3) find the codewords for the datawords (a) [0000]; (b) [1111]; and (c) [0010].

11.6. Using the parity check matrix of Eq. (11.5) find the syndrome for the codeword (1110100). Comment on this.

11.7. Using the parity check matrix of Eq. (11.5) find the syndrome for the codeword (1000110). Comment on this.

11.8. A received codeword is 1011000. Determine, using the parity check matrix of Eq. (11.5) if this is a valid codeword, and if not, write out the error vector on the assumption that only one error is present.

11.9. Calculate the code rate for a Hamming (31, 26) code.

11.10. Calculate the code rate and the number of errors that can be corrected with a BCH (63, 36) code.

11.11. An R-S code is byte oriented with $k = 8$. Given that there are eight redundant symbols, calculate the number of symbol errors that can be corrected.

11.12. Determine the values for N' and K' for the shortened R-S codes used in (*a*) DirecTV and (*b*) DVB.

11.13. Describe how convolution coding is achieved. State some of the main advantages and disadvantages of this type of code compared with block codes.

11.14. Explain what is meant by *interleaving* when applied to error control coding and why this might be used.

11.15. Explain what is meant by *concatenated codes* and why these might be used.

11.16. Explain what is meant by a FEC code. FEC coding at a code rate of 3/4 is used in a digital system. Given that the message bit rate is 1.544 Mb/s, calculate the transmission rate.

11.17. The bit rate for a baseband signal is 1.544 Mb/s, and FEC at a code rate of 7/8 is applied before the signal is used to modulate the carrier. Given that the system uses raised-cosine filtering with a rolloff factor of 0.2, determine the bandwidth required for (*a*) BPSK, and (*b*) QPSK.

11.18. A BPSK signal provides an $[E_b/N_0]$ of 9 dB at the receiver. Calculate the probability of bit error.

11.19. For the signal in Prob. 11.18, calculate the new value of bit error probability if FEC is applied at a code rate of 3/4, given that the carrier power remains unchanged.

11.20. Derive Eq. (11.11).

11.21. Explain what is meant by *coding gain* as applied to error correcting coding. When FEC coding is used on a digital link, a coding gain of 3 dB is achieved for the same BER as the uncoded case. What decibel reduction in transmitted carrier power does this imply?

11.22. A certain (15, 11) block code is capable of correcting one error at most. Given that this is a perfect code (see Sec. 11.8), plot, on the same set of axes, the BER for the coded and uncoded cases for an $[E_b/N_0]$ range of 2 to 12 dB. Calculate the coding gain at a BER of 10^{-6}.

11.23. State briefly the difference between hard and soft decision decoding. Following the description given in Sec. 11.9, determine the output produced by (*a*) hard decision and (*b*) soft decision decoding when the sampled signal from the demodulator is -0.4 V, 0.85 V, and -0.4 V for triple redundancy coding.

11.24. From Fig. 11.10 find the minimum $[E_b/N_0]$ as determined by the Shannon limit curve. Explain the significance of this.

11.25. For equiprobable bit transmission a received bit level is 0.55V. Assuming this is normalized where 1V represents a certainty of the bit being a 1, calculate the LLR.

11.26. Referring to Eq. (11.26), identify the parity equations which contain code bit c_7.

11.27. Complete the Tanner graph of Fig. 11.13 for all the parity equations given in Eq. 11.26.

11.28. For the parity equation for row 5, (Eq. 11.26) the probabilities are: $p_5^1 = 0.7$, $p_8^1 = 0.6$, $p_{14}^1 = 0.3$. Calculate the estimated probability p_2.

11.29. Research the literature and write brief comparative notes on the use of turbo codes and LDPC codes in satellite communications.

11.30. A (31, 6) block code is used in an ARQ scheme. Determine the upper bound on the probability of bit error.

References

Berrou, C., A. Glavieux, and P. Thitmajshima. 1993. "Near Shannon Limit error-correcting coding: Turbo codes." *Proc. of the IEEE Int. Conf. Commun.*, Switzerland, at www-elec.enst-bretagne.fr/equipe/berrou/Near%20Shannon%20Limit%20Error.pdf

Breynaert, D. 2005. Analysis of the bandwidth efficiency of DVB-S2 in a typical data distribution network. CCBN, Beijing, March, at www.newtecamerica.com/news/articles/DVB-S2%20whitepaper.pdf

Burr, A. 2001. "Turbo Codes: The Ultimate Error Control Codes?" *Electron. Commun. Eng. J.*, August, pp. 155–165.

Comtech, 2002. The Case for Turbo Product Coding in Satellite Communications, at www.linksite/manuals/whitepapers/Comtech%20EF%20Data%20Case%20for%20Turbo%20Product%20Coding.pdf

Gallagher, R. G. 1962. "Low Density Parity Check Codes." *IRE Trans. Info. Theory 8*, pp. 21–28.

Mead, D. C. 2000. *Direct Broadcast Satellite Communications*. Addison-Wesley, Reading, MA.

Pratt, Timothy, and C. W. Bostian. 1986. *Satellite Communications*. Wiley, New York.

Proakis, J. G., and M. Salehi. 1994. *Communication Systems Engineering*. Prentice-Hall, Englewood Cliffs, NJ.

Roddy, D., and J. Coolen. 1994. *Electronic Communications*. 4th ed. Prentice-Hall, Englewood Cliffs, NJ.

Shannon, C. E. 1948. "A Mathematical Theory of Communications." *BSTJ*, Vol. 27, pp. 379–423, 623–656, July, October.

Summers, T. 2004. "LDPC: Another Key Step toward Shannon." *CommsDesign*, October, at www.commsdesign.com/design_corner/showArticle.jhtml?articleID=49901136-50k

Talk Satellite 2004. Spectra Licensing Group Announces Gilat's Entrance into the Turbo Code Licensing Program for DVB-RCS Satellite Applications. December 15, at www.talksatellite.com/EMEAdoc105

Tanner, R. M., 1981. "A Recursive Approach to Low Complexity Codes." *IEEE Trans. Inf. Theory*, Vol. 27, No. 5, September, pp. 533–547.

Taub, H., and D. L. Schilling. 1986. *Principles of Communications Systems*, 2d ed. McGraw-Hill, New York.

Yoshida, J. 2003. "Hughes Goes Retro in Digital Satellite TV Coding." *EETimesUK*, November, at www.eetuk.com/tech/news/OEG20031110S0081-42k-

Chapter

12

The Space Link

12.1 Introduction

This chapter describes how the link-power budget calculations are made. These calculations basically relate two quantities, the transmit power and the receive power, and show in detail how the difference between these two powers is accounted for.

Link-budget calculations are usually made using decibel or decilog quantities. These are explained in App. G. In this text [square] brackets are used to denote decibel quantities using the basic power definition. Where no ambiguity arises regarding the units, the abbreviation dB is used. For example, Boltzmann's constant is given as −228.6 dB, although, strictly speaking, this should be given as −228.6 decilogs relative to 1 J/K. Where it is desirable to show the reference unit, this is indicated in the abbreviation, for example, dBHz means decibels relative to 1 Hz.

12.2 Equivalent Isotropic Radiated Power

A key parameter in link-budget calculations is the *equivalent isotropic radiated power*, conventionally denoted as EIRP. From Eqs. (6.4) and (6.5), the maximum power flux density at some distance r from a transmitting antenna of gain G is

$$\Psi_M = \frac{\mathrm{GP}_S}{4\pi r^2} \tag{12.1}$$

An isotropic radiator with an input power equal to GP_S would produce the same flux density. Hence, this product is referred to as the EIRP, or

$$\mathrm{EIRP} = \mathrm{GP}_S \tag{12.2}$$

EIRP is often expressed in decibels relative to 1 W, or dBW. Let P_S be in watts; then

$$[\text{EIRP}] = [P_S] + [G] \text{ dBW} \tag{12.3}$$

where $[P_S]$ is also in dBW and $[G]$ is in dB.

Example 12.1 A satellite downlink at 12 GHz operates with a transmit power of 6 W and an antenna gain of 48.2 dB. Calculate the EIRP in dBW.

Solution

$$[\text{EIRP}] = 10 \log\left(\frac{6\text{W}}{1\text{W}}\right) + 48.2$$

$$= \underline{\underline{56 \text{ dBW}}}$$

For a paraboloidal antenna, the isotropic power gain is given by Eq. (6.32). This equation may be rewritten in terms of frequency, since this is the quantity which is usually known.

$$G = \eta(10.472fD)^2 \tag{12.4}$$

where f is the carrier frequency in GHz, D is the reflector diameter in m, and η is the aperture efficiency. A typical value for aperture efficiency is 0.55, although values as high as 0.73 have been specified (Andrew Antenna, 1985).

With the diameter D in feet and all other quantities as before, the equation for power gain becomes

$$G = \eta(3.192fD)^2 \tag{12.5}$$

Example 12.2 Calculate the gain in decibels of a 3-m paraboloidal antenna operating at a frequency of 12 GHz. Assume an aperture efficiency of 0.55.

Solution

$$G = 0.55 \times (10.472 \times 12 \times 3)^2 \cong 78168$$

Hence,

$$[G] = 10 \log 78168 = 48.9 \text{ dB}$$

12.3 Transmission Losses

The [EIRP] may be thought of as the power input to one end of the transmission link, and the problem is to find the power received at the other end. Losses will occur along the way, some of which are constant.

Other losses can only be estimated from statistical data, and some of these are dependent on weather conditions, especially on rainfall.

The first step in the calculations is to determine the losses for *clear-weather* or *clear-sky conditions*. These calculations take into account the losses, including those calculated on a statistical basis, which do not vary significantly with time. Losses which are weather-related, and other losses which fluctuate with time, are then allowed for by introducing appropriate *fade margins* into the transmission equation.

12.3.1 Free-space transmission

As a first step in the loss calculations, the power loss resulting from the spreading of the signal in space must be determined. This calculation is similar for the uplink and the downlink of a satellite circuit. Using Eqs. (12.1) and (12.2) gives the power-flux density at the receiving antenna as

$$\Psi_M = \frac{\text{EIRP}}{4\pi r^2} \tag{12.6}$$

The power delivered to a matched receiver is this power-flux density multiplied by the effective aperture of the receiving antenna, given by Eq. (6.15). The received power is therefore

$$\begin{aligned} P_R &= \Psi_M A_{\text{eff}} \\ &= \frac{\text{EIRP}}{4\pi r^2}\frac{\lambda^2 G_R}{4\pi} \\ &= (\text{EIRP})(G_R)\left(\frac{\lambda}{4\pi r}\right)^2 \end{aligned} \tag{12.7}$$

Recall that r is the distance, or range, between the transmit and receive antennas and G_R is the isotropic power gain of the receiving antenna. The subscript R is used to identify the receiving antenna.

The right-hand side of Eq. (12.7) is separated into three terms associated with the transmitter, receiver, and free space, respectively. In decibel notation, the equation becomes

$$[P_R] = [\text{EIRP}] + [G_R] - 10 \log\left(\frac{4\pi r}{\lambda}\right)^2 \tag{12.8}$$

The received power in dBW is therefore given as the sum of the transmitted EIRP in dBW plus the receiver antenna gain in dB minus a third term, which represents the free-space loss in decibels. The free-space loss component in decibels is given by

$$[\text{FSL}] = 10 \log\left(\frac{4\pi r}{\lambda}\right)^2 \tag{12.9}$$

Normally, the frequency rather than wavelength will be known, and the substitution $\lambda = c/f$ can be made, where $c = 10^8$ m/s. With frequency in megahertz and distance in kilometers, it is left as an exercise for the student to show that the free-space loss is given by

$$[\text{FSL}] = 32.4 + 20 \log r + 20 \log f \qquad (12.10)$$

Equation (12.8) can then be written as

$$[P_R] = [\text{EIRP}] + [G_R] - [\text{FSL}] \qquad (12.11)$$

The received power $[P_R]$ will be in dBW when the [EIRP] is in dBW, and [FSL] in dB. Equation (12.9) is applicable to both the uplink and the downlink of a satellite circuit, as will be shown in more detail shortly.

Example 12.3 The range between a ground station and a satellite is 42,000 km. Calculate the free-space loss at a frequency of 6 GHz.

Solution

$$[\text{FSL}] = 32.4 + 20 \log 42{,}000 + 20 \log 6000 = \underline{\underline{200.4 \text{ dB}}}$$

This is a very large loss. Suppose that the [EIRP] is 56 dBW (as calculated in Example 12.1 for a radiated power of 6 W) and the receive antenna gain is 50 dB. The receive power would be 56 + 50 − 200.4 = −94.4 dBW. This is 355 pW. It also may be expressed as −64.4 dBm, which is 64.4 dB below the 1-mW reference level.

Equation (12.11) shows that the received power is increased by increasing antenna gain as expected, and Eq. (6.32) shows that antenna gain is inversely proportional to the square of the wavelength. Hence, it might be thought that increasing the frequency of operation (and therefore decreasing wavelength) would increase the received power. However, Eq. (12.9) shows that the free-space loss is also inversely proportional to the square of the wavelength, so these two effects cancel. It follows, therefore, that for a constant EIRP, the received power is independent of frequency of operation.

If the transmit power is a specified constant, rather than the EIRP, then the received power will increase with increasing frequency for given antenna dish sizes at the transmitter and receiver. It is left as an exercise for the student to show that under these conditions the received power is directly proportional to the square of the frequency.

12.3.2 Feeder losses

Losses will occur in the connection between the receive antenna and the receiver proper. Such losses will occur in the connecting waveguides, filters, and couplers. These will be denoted by RFL, or [RFL] dB, for *receiver*

feeder losses. The [RFL] values are added to [FSL] in Eq. (12.11). Similar losses will occur in the filters, couplers, and waveguides connecting the transmit antenna to the *high-power amplifier* (HPA) output. However, provided that the EIRP is stated, Eq. (12.11) can be used without knowing the transmitter feeder losses. These are needed only when it is desired to relate EIRP to the HPA output, as described in Secs. 12.7.4 and 12.8.2.

12.3.3 Antenna misalignment losses

When a satellite link is established, the ideal situation is to have the earth station and satellite antennas aligned for maximum gain, as shown in Fig. 12.1*a*. There are two possible sources of off-axis loss, one at the satellite and one at the earth station, as shown in Fig. 12.1*b*.

The off-axis loss at the satellite is taken into account by designing the link for operation on the actual satellite antenna contour; this is described in more detail in later sections. The off-axis loss at the earth station is referred to as the *antenna pointing loss*. Antenna pointing losses are usually only a few tenths of a decibel; typical values are given in Table 12.1.

In addition to pointing losses, losses may result at the antenna from misalignment of the polarization direction (these are in addition to the polarization losses described in Chap. 5). The polarization misalignment losses are usually small, and it will be assumed that the antenna misalignment losses, denoted by [AML], include both pointing and polarization losses resulting from antenna misalignment. It should be noted

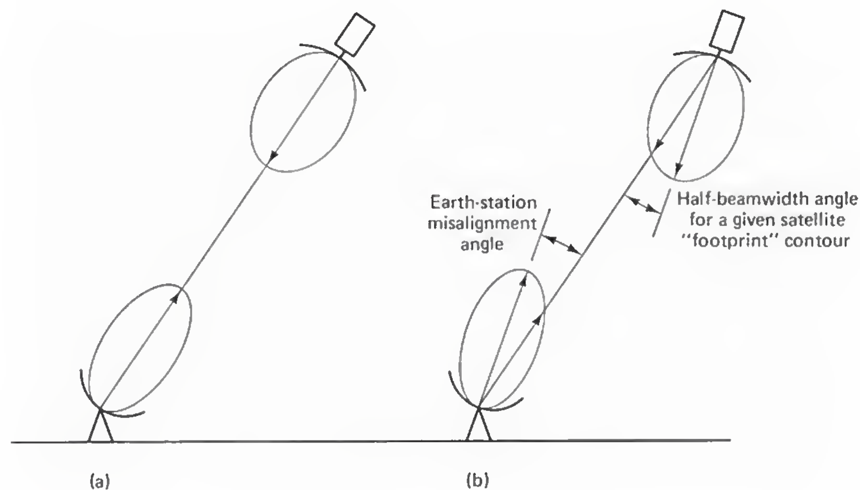


Figure 12.1 (*a*) Satellite and earth-station antennas aligned for maximum gain; (*b*) earth station situated on a given satellite "footprint," and earth-station antenna misaligned.

TABLE 12.1 Atmospheric Absorption Loss and Satellite Pointing Loss for Cities and Communities in the Province of Ontario

Location	Atmospheric absorption dB, summer	Satellite antenna pointing loss, dB	
		¼ Canada coverage	½ Canada coverage
Cat Lake	0.2	0.5	0.5
Fort Severn	0.2	0.9	0.9
Geraldton	0.2	0.2	0.1
Kingston	0.2	0.5	0.4
London	0.2	0.3	0.6
North Bay	0.2	0.3	0.2
Ogoki	0.2	0.4	0.3
Ottawa	0.2	0.6	0.2
Sault Ste. Marie	0.2	0.1	0.3
Sioux Lookout	0.2	0.4	0.3
Sudbury	0.2	0.3	0.2
Thunder Bay	0.2	0.3	0.2
Timmins	0.2	0.5	0.2
Toronto	0.2	0.3	0.4
Windsor	0.2	0.5	0.8

SOURCE: Telesat Canada Design Workbook.

that the antenna misalignment losses have to be estimated from statistical data, based on the errors actually observed for a large number of earth stations, and of course, the separate antenna misalignment losses for the uplink and the downlink must be taken into account.

12.3.4 Fixed atmospheric and ionospheric losses

Atmospheric gases result in losses by absorption, as described in Sec. 4.2 and by Eq. (4.1). These losses usually amount to a fraction of a decibel, and in subsequent calculations, the decibel value will be denoted by [AA]. Values obtained for some locations in the Province of Ontario, Canada, are shown in Table 12.1. Also, as discussed in Sec. 5.5, the ionosphere introduces a depolarization loss given by Eq. (5.19), and in subsequent calculations, the decibel value for this will be denoted by [PL].

12.4 The Link-Power Budget Equation

As mentioned at the beginning of Sec. 12.3, the [EIRP] can be considered as the input power to a transmission link. Now that the losses for the link have been identified, the power at the receiver, which is the power output of the link, may be calculated simply as [EIRP] − [LOSSES] + $[G_R]$, where the last quantity is the receiver antenna gain. Note carefully that decibel addition must be used.

The major source of loss in any ground-satellite link is the free-space spreading loss [FSL], as shown in Sec. 12.3.1, where Eq. (12.13) is the basic link-power budget equation taking into account this loss only. However, the other losses also must be taken into account, and these are simply added to [FSL]. The losses for clear-sky conditions are

$$[\text{LOSSES}] = [\text{FSL}] + [\text{RFL}] + [\text{AML}] + [\text{AA}] + [\text{PL}] \quad (12.12)$$

The decibel equation for the received power is then

$$[P_R] = [\text{EIRP}] + [G_R] - [\text{LOSSES}] \quad (12.13)$$

where [PR] = received power, dBW
[EIRP] = equivalent isotropic radiated power, dBW
[FSL] = free-space spreading loss, dB
[RFL] = receiver feeder loss, dB
[AML] = antenna misalignment loss, dB
[AA] = atmospheric absorption loss, dB
[PL] = polarization mismatch loss, dB

Example 12.4 A satellite link operating at 14 GHz has receiver feeder losses of 1.5 dB and a free-space loss of 207 dB. The atmospheric absorption loss is 0.5 dB, and the antenna pointing loss is 0.5 dB. Depolarization losses may be neglected. Calculate the total link loss for clear-sky conditions.

Solution The total link loss is the sum of all the losses:

$$\begin{aligned}[\text{LOSSES}] &= [\text{FSL}] + [\text{RFL}] + [\text{AA}] + [\text{AML}] \\ &= 207 + 1.5 + 0.5 + 0.5 \\ &= \underline{\underline{209.5\ \text{dB}}}\end{aligned}$$

12.5 System Noise

It is shown in Sec. 12.3 that the receiver power in a satellite link is very small, on the order of picowatts. This by itself would be no problem because amplification could be used to bring the signal strength up to an acceptable level. However, electrical noise is always present at the input, and unless the signal is significantly greater than the noise, amplification will be of no help because it will amplify signal and noise to the same extent. In fact, the situation will be worsened by the noise added by the amplifier.

The major source of electrical noise in equipment is that which arises from the random thermal motion of electrons in various resistive and active devices in the receiver. Thermal noise is also generated in the

lossy components of antennas, and thermal-like noise is picked up by the antennas as radiation.

The available noise power from a thermal noise source is given by

$$P_N = kT_N B_N \tag{12.14}$$

Here, T_N is known as the equivalent noise temperature, B_N is the equivalent noise bandwidth, and $k = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant. With the temperature in kelvins and bandwidth in hertz, the noise power will be in watts. The noise power bandwidth is always wider than the −3-dB bandwidth determined from the amplitude-frequency response curve, and a useful rule of thumb is that the noise bandwidth is equal to 1.12 times the −3-dB bandwidth, or $B_N \approx 1.12 \times B_{-3\text{dB}}$. The bandwidths here are in hertz (or a multiple such as MHz).

The main characteristic of thermal noise is that it has a *flat frequency spectrum*; that is, the noise power per unit bandwidth is a constant. The noise power per unit bandwidth is termed the *noise power spectral density*. Denoting this by N_0, then from Eq. (12.14),

$$N_0 = \frac{P_N}{B_N} = kT_N \text{ J} \tag{12.15}$$

The noise temperature is directly related to the physical temperature of the noise source but is not always equal to it. This is discussed more fully in the following sections. The noise temperatures of various sources which are connected together can be added directly to give the total noise.

Example 12.5 An antenna has a noise temperature of 35 K and is matched into a receiver which has a noise temperature of 100 K. Calculate (*a*) the noise power density and (*b*) the noise power for a bandwidth of 36 MHz.

Solution

$$(a)\ N_0 = (35 + 100) \times 1.38 \times 10^{-23} = \underline{\underline{1.86}} \times \underline{\underline{10^{-21}\text{ J}}}$$

$$(b)\ P_N = 1.86 \times 10^{-21} \times 36 \times 10^6 = \underline{\underline{0.067\text{ pW}}}$$

In addition to these thermal noise sources, intermodulation distortion in high-power amplifiers (see Sec. 12.7.3) can result in signal products which appear as noise and in fact is referred to as *intermodulation noise*. This is discussed in Sec. 12.10.

12.5.1 Antenna noise

Antennas operating in the receiving mode introduce noise into the satellite circuit. Noise therefore will be introduced by the satellite receive antenna and the ground station receive antenna. Although the physical

origins of the noise in either case are similar, the magnitudes of the effects differ significantly.

The antenna noise can be broadly classified into two groups: noise originating from antenna losses and *sky noise.* Sky noise is a term used to describe the microwave radiation which is present throughout the universe and which appears to originate from matter in any form at finite temperatures. Such radiation in fact covers a wider spectrum than just the microwave spectrum. The equivalent noise temperature of the sky, as seen by an earth-station antenna, is shown in Fig. 12.2. The lower graph is for the antenna pointing directly overhead, while the upper graph is for the antenna pointing just above the horizon. The increased noise in the latter case results from the thermal radiation of the earth, and this in fact sets a lower limit of about 5° at C band and 10° at Ku band on the elevation angle which may be used with ground-based antennas.

The graphs show that at the low-frequency end of the spectrum, the noise decreases with increasing frequency. Where the antenna is zenith-pointing, the noise temperature falls to about 3 K at frequencies between

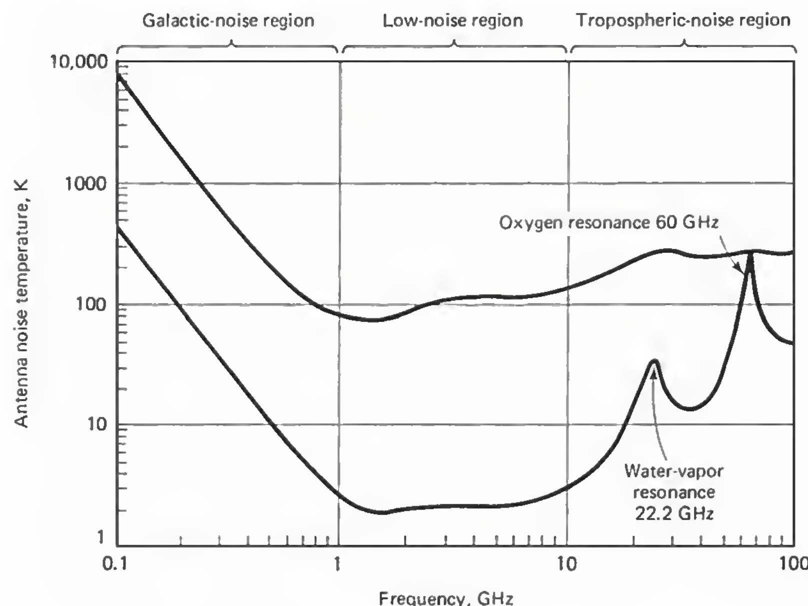


Figure 12.2 Irreducible noise temperature of an ideal, ground-based antenna. The antenna is assumed to have a very narrow beam without sidelobes or electrical losses. Below 1 GHz, the maximum values are for the beam pointed at the galactic poles. At higher frequencies, the maximum values are for the beam just above the horizon and the minimum values for zenith pointing. The low-noise region between 1 and 10 GHz is most amenable to application of special, low-noise antennas. (*From Philip F. Panter, "Communications Systems Design," McGraw-Hill Book Company, New York, 1972. With permission.*)

about 1 and 10 GHz. This represents the residual background radiation in the universe. Above about 10 GHz, two peaks in temperature are observed, resulting from resonant losses in the earth's atmosphere. These are seen to coincide with the peaks in atmospheric absorption loss shown in Fig. 4.2.

Any absorptive loss mechanism generates thermal noise, there being a direct connection between the loss and the effective noise temperature, as shown in Sec. 12.5.5. Rainfall introduces attenuation, and therefore, it degrades transmissions in two ways: It attenuates the signal, and it introduces noise. The detrimental effects of rain are much worse at Ku-band frequencies than at C band, and the downlink rain-fade margin, discussed in Sec. 12.9.2, must also allow for the increased noise generated.

Figure 12.2 applies to ground-based antennas. Satellite antennas are generally pointed toward the earth, and therefore, they receive the full thermal radiation from it. In this case the equivalent noise temperature of the antenna, excluding antenna losses, is approximately 290 K.

Antenna losses add to the noise received as radiation, and the total antenna noise temperature is the sum of the equivalent noise temperatures of all these sources. For large ground-based C-band antennas, the total antenna noise temperature is typically about 60 K, and for the Ku band, about 80 K under clear-sky conditions. These values do not apply to any specific situation and are quoted merely to give some idea of the magnitudes involved. Figure 12.3 shows the noise temperature as a function of angle of elevation for a 1.8-m antenna operating in the Ku band.

12.5.2 Amplifier noise temperature

Consider first the noise representation of the antenna and the *low noise amplifier* (LNA) shown in Fig. 12.4*a*. The available power gain of the amplifier is denoted as G, and the noise power output, as P_{no}. For the

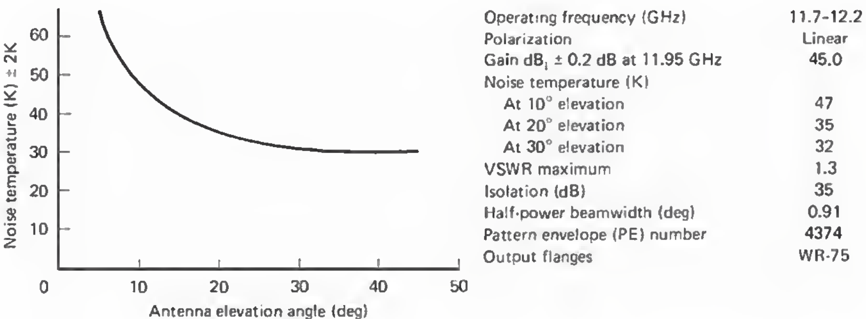


Figure 12.3 Antenna noise temperature as a function of elevation for 1.8-m antenna characteristics. (*Andrew Bulletin 1206; courtesy of Andrew Antenna Company, Limited.*)

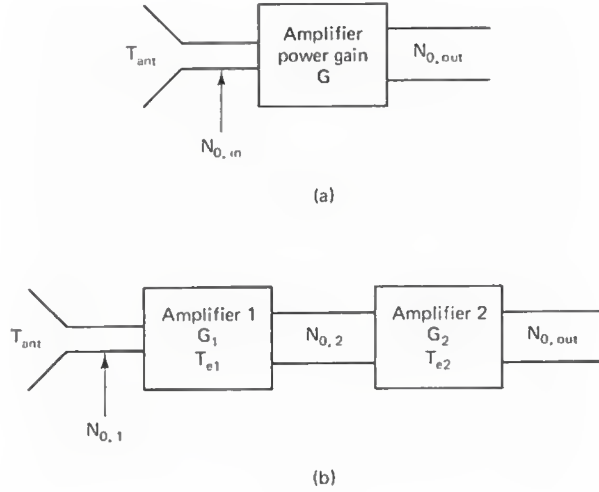


Figure 12.4 Circuit used in finding equivalent noise temperature of (*a*) an amplifier and (*b*) two amplifiers in cascade.

moment we will work with the noise power per unit bandwidth, which is simply noise energy in joules as shown by Eq. (12.15). The input noise energy coming from the antenna is

$$N_{0,\mathrm{ant}} = kT_{\mathrm{ant}} \tag{12.16}$$

The output noise energy $N_{0,\mathrm{out}}$ will be $GN_{0,\mathrm{ant}}$ plus the contribution made by the amplifier. Now all the amplifier noise, wherever it occurs in the amplifier, may be *referred to the input* in terms of an equivalent input noise temperature for the amplifier T_e. This allows the output noise to be written as

$$N_{0,\mathrm{out}} = Gk(T_{\mathrm{ant}} + T_e) \tag{12.17}$$

The total noise referred to the input is simply $N_{0,\mathrm{out}}/G$, or

$$N_{0,\mathrm{in}} = k(T_{\mathrm{ant}} + T_e) \tag{12.18}$$

T_e can be obtained by measurement, a typical value being in the range 35 to 100 K. Typical values for T_{ant} are given in Sec. 12.5.1.

12.5.3 Amplifiers in cascade

The cascade connection is shown in Fig. 12.4*b*. For this arrangement, the overall gain is

$$G = G_1G_2 \tag{12.19}$$

The noise energy of amplifier 2 referred to its own input is simply kT_{e2}. The noise input to amplifier 2 from the preceding stages is $G_1k(T_{ant} + T_{e1})$, and thus the total noise energy *referred to amplifier 2 input* is

$$N_{0,2} = G_1k(T_{ant} + T_{e1}) + kT_{e2} \tag{12.20}$$

This noise energy may be referred to amplifier 1 input by dividing by the available power gain of amplifier 1:

$$\begin{aligned} N_{0,1} &= \frac{N_{0,2}}{G_1} \\ &= k\left(T_{ant} + T_{e1} + \frac{T_{e2}}{G_1}\right) \end{aligned} \tag{12.21}$$

A system noise temperature may now be defined as T_S by

$$N_{0,1} = kT_S \tag{12.22}$$

and hence it will be seen that T_S is given by

$$T_S = T_{ant} + T_{e1} + \frac{T_{e2}}{G_1} \tag{12.23}$$

This is a very important result. It shows that the noise temperature of the second stage is divided by the power gain of the first stage when referred to the input. Therefore, in order to keep the overall system noise as low as possible, the first stage (usually an LNA) should have high power gain as well as low noise temperature.

This result may be generalized to any number of stages in cascade, giving

$$T_S = T_{ant} + T_{e1} + \frac{T_{e2}}{G_1} + \frac{T_{e3}}{G_1G_2} + \cdots \tag{12.24}$$

12.5.4 Noise factor

An alternative way of representing amplifier noise is by means of its *noise factor, F*. In defining the noise factor of an amplifier, the source is taken to be at *room temperature*, denoted by T_0, usually taken as 290 K. The input noise from such a source is kT_0, and the output noise from the amplifier is

$$N_{0,out} = FGkT_0 \tag{12.25}$$

Here, G is the available power gain of the amplifier as before, and F is its noise factor.

A simple relationship between noise temperature and noise factor can be derived. Let T_e be the noise temperature of the amplifier, and let the source be at room temperature as required by the definition of F. This means that $T_{ant} = T_0$. Since the same noise output must be available whatever the representation, it follows that

$$Gk(T_0 + T_e) = FGkT_0$$

or

$$T_e = (F - 1)\ T_0 \tag{12.26}$$

This shows the direct equivalence between noise factor and noise temperature. As a matter of convenience, in a practical satellite receiving system, noise temperature is specified for low-noise amplifiers and converters, while noise factor is specified for the main receiver unit.

The *noise figure* is simply F expressed in decibels:

$$\text{Noise figure} = [F] = 10 \log F \tag{12.27}$$

Example 12.6 An LNA is connected to a receiver which has a noise figure of 12 dB. The gain of the LNA is 40 dB, and its noise temperature is 120 K. Calculate the overall noise temperature referred to the LNA input.

Solution 12 dB is a power ratio of 15.85:1, and therefore,

$$T_{e2} = (15.85 - 1) \times 290 = 4306 \text{ K}$$

A gain of 40 dB is a power ratio of 10^4:1, and therefore,

$$T_{in} = 120 + \frac{4306}{10^4}$$

$$= \underline{\underline{120.43 \text{ K}}}$$

In Example 12.6 it will be seen that the decibel quantities must be converted to power ratios. Also, even though the main receiver has a very high noise temperature, its effect is made negligible by the high gain of the LNA.

12.5.5 Noise temperature of absorptive networks

An *absorptive network* is one which contains resistive elements. These introduce losses by absorbing energy from the signal and converting it to heat. Resistive attenuators, transmission lines, and waveguides are all examples of absorptive networks, and even rainfall, which absorbs energy from radio signals passing through it, can be considered a form

of absorptive network. Because an absorptive network contains resistance, it generates thermal noise.

Consider an absorptive network, which has a power loss L. The power loss is simply the ratio of input power to output power and will always be greater than unity. Let the network be matched at both ends, to a terminating resistor, R_T, at one end and an antenna at the other, as shown in Fig. 12.5, and let the system be at some ambient temperature T_x. The noise energy transferred from R_T into the network is kT_x. Let the network noise be represented at the output terminals (the terminals connected to the antenna in this instance) by an equivalent noise temperature $T_{\mathrm{NW},0}$. Then the noise energy radiated by the antenna is

$$N_{\mathrm{rad}} = \frac{kT_x}{L} + kT_{\mathrm{NW},0} \tag{12.28}$$

Because the antenna is matched to a resistive source at temperature T_x, the available noise energy which is fed into the antenna and radiated is $N_{\mathrm{rad}} = kT_x$. Keep in mind that the antenna resistance to which the network is matched is fictitious, in the sense that it represents radiated power, but it does not generate noise power. This expression for N_{rad} can be substituted into Eq. (12.28) to give

$$T_{\mathrm{NW},0} = T_x\left(1 - \frac{1}{L}\right) \tag{12.29}$$

This is the equivalent noise temperature of the network referred to the output terminals of the network. The equivalent noise at the output can be transferred to the input on dividing by the network power gain, which by definition is $1/L$. Thus, the equivalent noise temperature of the network referred to the network input is

$$T_{\mathrm{NW},i} = T_x(L - 1) \tag{12.30}$$

Since the network is bilateral, Eqs. (12.29) and (12.30) apply for signal flow in either direction. Thus, Eq. (12.30) gives the equivalent noise

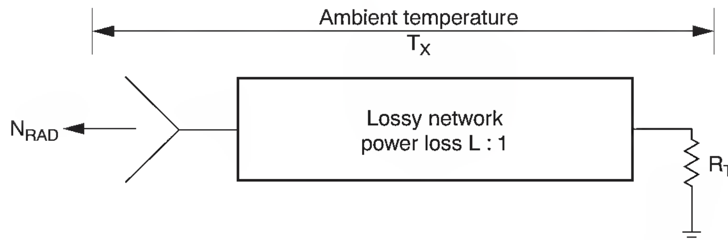


Figure 12.5 Network matched at both ends, to a terminating resistor R_T at one end and an antenna at the other.

temperature of a lossy network referred to the input at the antenna when the antenna is used in receiving mode.

If the lossy network should happen to be at room temperature, that is, $T_x = T_0$, then a comparison of Eqs. (12.26) and (12.30) shows that

$$F = L \tag{12.31}$$

This shows that at room temperature the noise factor of a lossy network is equal to its power loss.

12.5.6 Overall system noise temperature

Figure 12.6*a* shows a typical receiving system. Applying the results of the previous sections yields, for the system noise temperature referred to the input,

$$T_S = T_{\text{ant}} + T_{e1} + \frac{(L-1)T_0}{G_1} + \frac{L(F-1)T_0}{G_1} \tag{12.32}$$

The significance of the individual terms is illustrated in the following examples.

Example 12.7 For the system shown in Fig. 12.6*a*, the receiver noise figure is 12 dB, the cable loss is 5 dB, the LNA gain is 50 dB, and its noise temperature 150 K. The antenna noise temperature is 35 K. Calculate the noise temperature referred to the input.

Solution For the main receiver, $F = 10^{1.2} = 15.85$. For the cable, $L = 10^{0.5} = 3.16$. For the LNA, $G = 10^5$. Hence,

$$T_S = 35 + 150 + \frac{(3.16-1)\times 290}{10^5} + \frac{3.16\times(15.85-1)\times 290}{10^5}$$

$$\cong \underline{\underline{185\text{ K}}}$$

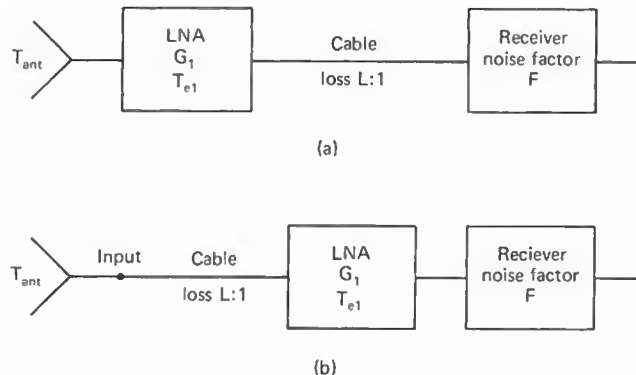


Figure 12.6 Connections used in examples illustrating overall noise temperature of system, Sec. 12.5.6.

Example 12.8 Repeat the calculation when the system of Fig. 12.6*a* is arranged as shown in Fig. 12.6*b*.

Solution In this case the cable precedes the LNA, and therefore, the equivalent noise temperature referred to the cable input is

$$T_S = 35 + (3.16 - 1) \times 290 + 3.16 \times 150 + \frac{3.16 \times (15.85 - 1) \times 290}{10^5}$$

$$= \underline{\underline{1136 \text{ K}}}$$

Examples 12.7 and 12.8 illustrate the important point that the LNA must be placed ahead of the cable, which is why one sees amplifiers mounted right at the dish in satellite receive systems.

12.6 Carrier-to-Noise Ratio

A measure of the performance of a satellite link is the ratio of carrier power to noise power at the receiver input, and link-budget calculations are often concerned with determining this ratio. Conventionally, the ratio is denoted by C/N (or CNR), which is equivalent to P_R/P_N. In terms of decibels,

$$\left[\frac{C}{N}\right] = [P_R] - [P_N] \qquad (12.33)$$

Equations (12.17) and (12.18) may be used for $[P_R]$ and $[P_N]$, resulting in

$$\left[\frac{C}{N}\right] = [\text{EIRP}] + [G_R] - [\text{LOSSES}] - [k] - [T_S] - [B_N] \qquad (12.34)$$

The G/T ratio is a key parameter in specifying the receiving system performance. The antenna gain G_R and the system noise temperature T_S can be combined in Eq. (12.34) as

$$[G/T] = [G_R] - [T_S] \text{ dBK}^{-1} \qquad (12.35)$$

Therefore, the link equation [Eq. (12.34)] becomes

$$\left[\frac{C}{N}\right] = [\text{EIRP}] + \left[\frac{G}{T}\right] - [\text{LOSSES}] - [k] - [B_N] \qquad (12.36)$$

The ratio of carrier power to noise power density P_R/N_0 may be the quantity actually required. Since $P_N = kT_NB_N = N_0\text{B}_\text{N}$, then

$$\left[\frac{C}{N}\right] = \left[\frac{C}{N_0B_N}\right]$$

$$= \left[\frac{C}{N_0}\right] - [B_N]$$

and therefore

$$\left[\frac{C}{N_0}\right] = \left[\frac{C}{N}\right] + [B_N] \tag{12.37}$$

$[C/N]$ is a true power ratio in units of decibels, and $[B_N]$ is in decibels relative to 1 Hz, or dBHz. Thus, the units for $[C/N_0]$ are dBHz.

Substituting Eq. (12.37) for $[C/N]$ gives

$$\left[\frac{C}{N_0}\right] = [\text{EIRP}] + \left[\frac{G}{T}\right] - [\text{LOSSES}] - [k] \tag{12.38}$$

Example 12.9 In a link-budget calculation at 12 GHz, the free-space loss is 206 dB, the antenna pointing loss is 1 dB, and the atmospheric absorption is 2 dB. The receiver $[G/T]$ is 19.5 dB/K, and receiver feeder losses are 1 dB. The EIRP is 48 dBW. Calculate the carrier-to-noise spectral density ratio.

Solution The data are best presented in tabular form and in fact lend themselves readily to spreadsheet-type computations. For brevity, the units are shown as decilogs, (see App. G) and losses are entered as negative numbers to take account of the minus sign in Eq. (12.38). Recall that Boltzmann's constant equates to −228.6 decilogs, so $-[k]$ = 228.6 decilogs, as shown in the following table.

Entering data in this way allows the final result to be entered in a table cell as the sum of the terms in the rows above the cell, a feature usually incorporated in spreadsheets and word processors. This is illustrated in the following table.

Quantity	Decilogs
Free-space loss	−206
Atmospheric absorption loss	−2
Antenna pointing loss	−1
Receiver feeder losses	−1
Polarization mismatch loss	0
Receiver G/T ratio	19.5
EIRP	48
$-[k]$	228.6
$[C/N_0]$, Eq. (12.38)	86.1

The final result, 86.10 dBHz, is the algebraic sum of the quantities as given in Eq. (12.38).

12.7 The Uplink

The uplink of a satellite circuit is the one in which the earth station is transmitting the signal and the satellite is receiving it. Equation (12.38) can be applied to the uplink, but subscript U will be used to denote

specifically that the uplink is being considered. Thus Eq. (12.38) becomes

$$\left[\frac{C}{N_0}\right]_U = [\text{EIRP}]_U + \left[\frac{G}{T}\right]_U - [\text{LOSSES}]_U - [k] \tag{12.39}$$

In Eq. (12.39) the values to be used are the earth station EIRP, the satellite receiver feeder losses, and satellite receiver G/T. The free-space loss and other losses which are frequency-dependent are calculated for the uplink frequency. The resulting carrier-to-noise density ratio given by Eq. (12.39) is that which appears at the satellite receiver.

In some situations, the flux density appearing at the satellite receive antenna is specified rather than the earth-station EIRP, and Eq. (12.39) is modified as explained next.

12.7.1 Saturation flux density

As explained in Sec. 7.7.3, the *traveling-wave tube amplifier* (TWTA) in a satellite transponder exhibits power output saturation, as shown in Fig. 7.21. The flux density required at the receiving antenna to produce saturation of the TWTA is termed the *saturation flux density*. The saturation flux density is a specified quantity in link budget calculations, and knowing it, one can calculate the required EIRP at the earth station. To show this, consider again Eq. (12.6) which gives the flux density in terms of EIRP, repeated here for convenience:

$$\Psi_M = \frac{\text{EIRP}}{4\pi r^2}$$

In decibel notation this is

$$[\Psi_M] = [\text{EIRP}] + 10\log\frac{1}{4\pi r^2} \tag{12.40}$$

But from Eq. (12.9) for free-space loss we have

$$-[\text{FSL}] = 10\log\frac{\lambda^2}{4\pi} + 10\log\frac{1}{4\pi r^2} \tag{12.41}$$

Substituting this in Eq. (12.40) gives

$$[\Psi_M] = [\text{EIRP}] - [\text{FSL}] - 10\log\frac{\lambda^2}{4\pi} \tag{12.42}$$

The $\lambda^2/4\pi$ term has dimensions of area, and in fact, from Eq. (6.15) it is the effective area of an isotropic antenna. Denoting this by A_0 gives

$$[A_0] = 10\log\frac{\lambda^2}{4\pi} \tag{12.43}$$

Since frequency rather than wavelength is normally known, it is left as an exercise for the student to show that with frequency f in gigahertz, Eq. (12.43) can be rewritten as

$$[A_0] = -(21.45 + 20 \log f) \tag{12.44}$$

Combining this with Eq. (12.42) and rearranging slightly gives the EIRP as

$$[\text{EIRP}] = [\Psi_M] + [A_0] + [\text{FSL}] \tag{12.45}$$

Equation (12.45) was derived on the basis that the only loss present was the spreading loss, denoted by [FSL]. But, as shown in the previous sections, the other propagation losses are the atmospheric absorption loss, the polarization mismatch loss, and the antenna misalignment loss. When allowance is made for these, Eq. (12.45) becomes

$$[\text{EIRP}] = [\Psi_M] + [A_0] + [\text{FSL}] + [\text{AA}] + [\text{PL}] + [\text{AML}] \tag{12.46}$$

In terms of the total losses given by Eq. (12.12), Eq. (12.46) becomes

$$[\text{EIRP}] = [\Psi_M] + [A_0] + [\text{LOSSES}] - [\text{RFL}] \tag{12.47}$$

This is for clear-sky conditions and gives the *minimum* value of [EIRP] which the earth station must provide to produce a given flux density at the satellite. Normally, the saturation flux density will be specified. With saturation values denoted by the subscript S, Eq. (12.47) is rewritten as

$$[\text{EIRP}_S]_U = [\Psi_S] + [A_0] + [\text{LOSSES}]_U - [\text{RFL}] \tag{12.48}$$

Example 12.10 An uplink operates at 14 GHz, and the flux density required to saturate the transponder is -120 dB(W/m^2). The free-space loss is 207 dB, and the other propagation losses amount to 2 dB. Calculate the earth-station [EIRP] required for saturation, assuming clear-sky conditions. Assume [RFL] is negligible.

Solution At 14 GHz,

$$[A_0] = -(21.45 + 20 \log 14) = -44.37 \text{ dB}$$

The losses in the propagation path amount to 207 + 2 = 209 dB. Hence, from Eq. (12.48),

$$[\text{EIRP}_S]_U = -120 - 44.37 + 209$$
$$= \underline{\underline{44.63 \text{ dBW}}}$$

12.7.2 Input backoff

As described in Sec. 12.7.3, where a number of carriers are present simultaneously in a TWTA, the operating point must be backed off to a linear portion of the transfer characteristic to reduce the effects of intermodulation distortion. Such multiple carrier operation occurs with *frequency-division multiple access* (FDMA), which is described in Chap. 14. The point to be made here is that *backoff* (BO) must be allowed for in the link-budget calculations.

Suppose that the saturation flux density for single-carrier operation is known. Input BO will be specified for multiple-carrier operation, referred to the single-carrier saturation level. The earth-station EIRP will have to be reduced by the specified BO, resulting in an uplink value of

$$[\mathrm{EIRP}]_U = [\mathrm{EIRP}_S]_U - [\mathrm{BO}]_i \tag{12.49}$$

Although some control of the input to the transponder power amplifier is possible through the ground TT&C station, as described in Sec. 12.7.3, input BO is normally achieved through reduction of the [EIRP] of the earth stations actually accessing the transponder.

Equations (12.48) and (12.49) may now be substituted in Eq. (12.39) to give

$$\left[\frac{C}{N_0}\right]_U = [\Psi_S] + [A_0] - [\mathrm{BO}]_i + \left[\frac{G}{T}\right]_U - [k] - [\mathrm{RFL}] \tag{12.50}$$

Example 12.11 An uplink at 14 GHz requires a saturation flux density of -91.4 dBW/m^2 and an input BO of 11 dB. The satellite [G/T] is -6.7 dBK^{-1}, and receiver feeder losses amount to 0.6 dB. Calculate the carrier-to-noise density ratio.

Solution As in Example 12.9, the calculations are best carried out in tabular form.

$[A_0] = -44.37$ dBm^2 for a frequency of 14 GHz is calculated by using Eq. (12.44) as in Example 12.10.

Quantity	Decilogs
Saturation flux density	−91.4
$[A_0]$ at 14 GHz	−44.4
Input BO	−11.0
Satellite saturation [G/T]	−6.7
$-[k]$	228.6
Receiver feeder loss	−0.6
Total	74.5

Note that $[k] = -228.6$ dB, so $-[k]$ in Eq. (12.50) becomes 228.6 dB. Also, [RFL] and $[\mathrm{BO}]_i$ are entered as negative numbers to take account of the minus signs attached to them in Eq. (12.50). The total gives the carrier-to-noise density ratio at the satellite receiver as 74.5 dBHz.

Since fade margins have not been included at this stage, Eq. (12.50) applies for *clear-sky* conditions. Usually, the most serious fading is caused by rainfall, as described in Sec. 12.9.

12.7.3 The earth station HPA

The earth station HPA has to supply the radiated power plus the transmit feeder losses, denoted here by TFL, or [TFL] dB. These include waveguide, filter, and coupler losses between the HPA output and the transmit antenna. Referring back to Eq. (12.3), the power output of the HPA is given by

$$[P_{\mathrm{HPA}}] = [\mathrm{EIRP}] - [G_T] + [\mathrm{TFL}] \tag{12.51}$$

The [EIRP] is that given by Eq. (12.49) and thus includes any input BO that is required at the satellite.

The earth station itself may have to transmit multiple carriers, and its output also will require back off, denoted by $[\mathrm{BO}]_{\mathrm{HPA}}$. The earth station HPA must be rated for a saturation power output given by

$$[P_{\mathrm{HPA,sat}}] = [P_{\mathrm{HPA}}] + [\mathrm{BO}]_{\mathrm{HPA}} \tag{12.52}$$

Of course, the HPA will be operated at the backed-off power level so that it provides the required power output $[P_{\mathrm{HPA}}]$. To ensure operation well into the linear region, an HPA with a comparatively high saturation level can be used and a high degree of BO introduced. The large physical size and high power consumption associated with larger tubes do not carry the same penalties they would if used aboard the satellite. Again, it is emphasized that BO at the earth station may be required quite independently of any BO requirements at the satellite transponder. The power rating of the earth-station HPA should also be sufficient to provide a fade margin, as discussed in Sec. 12.9.1.

12.8 Downlink

The downlink of a satellite circuit is the one in which the satellite is transmitting the signal and the earth station is receiving it. Equation (12.38) can be applied to the downlink, but subscript D will be used to denote specifically that the downlink is being considered. Thus Eq. (12.38) becomes

$$\left[\frac{C}{N_0}\right]_D = [\mathrm{EIRP}]_D + \left[\frac{G}{T}\right]_D - [\mathrm{LOSSES}]_D - [k] \tag{12.53}$$

In Eq. (12.53) the values to be used are the satellite EIRP, the earth-station receiver feeder losses, and the earth-station receiver G/T. The free space and other losses are calculated for the downlink frequency. The resulting carrier-to-noise density ratio given by Eq. (12.53) is that which appears at the detector of the earth station receiver.

Where the carrier-to-noise ratio is the specified quantity rather than carrier-to-noise density ratio, Eq. (12.38) is used. This becomes, on assuming that the signal bandwidth B is equal to the noise bandwidth B_N:

$$\left[\frac{C}{N}\right]_D = [\mathrm{EIRP}]_D + \left[\frac{G}{T}\right]_D - [\mathrm{LOSSES}]_D - [k] - [B] \qquad (12.54)$$

Example 12.12 A satellite TV signal occupies the full transponder bandwidth of 36 MHz, and it must provide a C/N ratio at the destination earth station of 22 dB. Given that the total transmission losses are 200 dB and the destination earth-station G/T ratio is 31 dB/K, calculate the satellite EIRP required.

Solution Equation (12.54) can be rearranged as

$$[\mathrm{EIRP}]_D = \left[\frac{C}{N}\right]_D - \left[\frac{G}{T}\right]_D + [\mathrm{LOSSES}]_D + [k] + [B]$$

Setting this up in tabular form, and keeping in mind that $+[k] = -228.6$ dB and that losses are numerically equal to $+200$ dB, we obtain

Quantity	Decilogs
$[C/N]$	22
$-[G/T]$	−31
[LOSSES]	200
$[k]$	−228.6
$[B]$	75.6
[EIRP]	38

The required EIRP is 38 dBW or, equivalently, 6.3 kW.

Example 12.12 illustrates the use of Eq. (12.54). Example 12.13 shows the use of Eq. (12.53) applied to a digital link.

Example 12.13 A QPSK signal is transmitted by satellite. Raised-cosine filtering is used, for which the rolloff factor is 0.2 and a *bit error rate* (BER) of 10^{-5} is required. For the satellite downlink, the losses amount to 200 dB, the receiving earth-station G/T ratio is 32 dBK^{-1}, and the transponder bandwidth is 36 MHz. Calculate (*a*) the bit rate which can be accommodated, and (*b*) the EIRP required.

Solution Equation (10.16) gives

$$R_b = \frac{2B}{1+\rho}$$
$$= \frac{2 \times 36 \times 10^6}{1.2}$$
$$= 60 \text{ Mbps}$$

Hence,

$$[R_b] = 10 \log\left(\frac{60 \times 10^6}{1\text{s}^{-1}}\right)$$
$$= 77.78 \text{ dBbps}$$

For BER = 10^{-5}, Fig. 10.17 gives an $[E_b/N_0]$ = 9.6 dB.

From Eq. (10.24) the required C/N_0 ratio is

$$\left[\frac{C}{N_0}\right] = \left[\frac{E_b}{N_0}\right] + [R_b]$$
$$= 77.78 + 9.6$$
$$= 87.38 \text{ dBHz}$$

From Eq. (12.53),

$$[\text{EIRP}]_D = \left[\frac{C}{N_0}\right]_D - \left[\frac{G}{T}\right]_D + [\text{LOSSES}]_D + [k]$$
$$= 87.38 - 32 + 200 - 228.6$$
$$\cong \underline{\underline{26.8 \text{ dBW}}}$$

12.8.1 Output back-off

Where input BO is employed as described in Sec. 12.7.2, a corresponding output BO must be allowed for in the satellite EIRP. As the curve of Fig. 7.21 shows, output BO is not linearly related to input BO. A rule of thumb, frequently used, is to take the output BO as the point on the curve which is 5 dB below the extrapolated linear portion, as shown in Fig. 12.7. Since the linear portion gives a 1:1 change in decibels, the relationship between input and output BO is $[\text{BO}]_0 = [\text{BO}]_i - 5$ dB. For example, with an input BO of $[\text{BO}]_i = 11$ dB, the corresponding output BO is $[\text{BO}]_0 = 11 - 5 = 6$ dB.

If the satellite EIRP for saturation conditions is specified as $[\text{EIRP}_S]_D$, then $[\text{EIRP}]_D = [\text{EIRP}_S]_D - [\text{BO}]_0$ and Eq. (12.53) becomes

$$\left[\frac{C}{N_0}\right]_D = [\text{EIRP}_S]_D - [\text{BO}]_0 + \left[\frac{G}{T}\right]_D - [\text{LOSSES}]_D - [k] \tag{12.55}$$

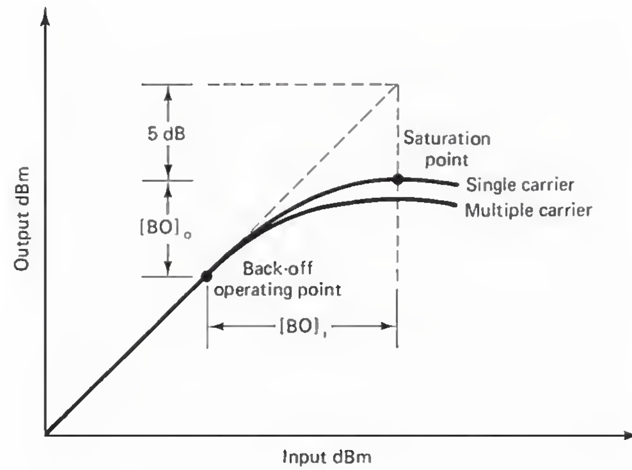


Figure 12.7 Input and output back-off relationship for the satellite traveling-wave-tube amplifier; $[\mathrm{BO}]_i = [\mathrm{BO}]_0 + 5$ dB.

Example 12.14 The specified parameters for a downlink are satellite saturation value of EIRP, 25 dBW; output BO, 6 dB; free-space loss, 196 dB; allowance for other downlink losses, 1.5 dB; and earth-station G/T, 41 dBK^{-1}. Calculate the carrier-to-noise density ratio at the earth station.

Solution As with the uplink budget calculations, the work is best set out in tabular form with the minus signs in Eq. (12.55) attached to the tabulated values.

Quantity	Decilogs
Satellite saturation [EIRP]	25.0
Free-space loss	−196.0
Other losses	−1.5
Output BO	−6.0
Earth station [G/T]	41.0
−[k]	228.6
Total	91.1

The total gives the carrier-to-noise density ratio at the earth station in dBHz, as calculated from Eq. (12.55).

For the uplink, the saturation flux density at the satellite receiver is a specified quantity. For the downlink, there is no need to know the saturation flux density at the earth-station receiver, since this is a terminal point, and the signal is not used to saturate a power amplifier.

12.8.2 Satellite TWTA output

The satellite power amplifier, which usually is a TWTA, has to supply the radiated power plus the transmit feeder losses. These losses include the waveguide, filter, and coupler losses between the TWTA output and

the satellite's transmit antenna. Referring back to Eq. (12.3), the power output of the TWTA is given by

$$[P_{\mathrm{TWTA}}] = [\mathrm{EIRP}]_D - [G_T]_D + [\mathrm{TFL}]_D \tag{12.56}$$

Once $[P_{\mathrm{TWTA}}]$ is found, the saturated power output rating of the TWTA is given by

$$[P_{\mathrm{TWTA}}]_S = [P_{\mathrm{TWTA}}] + [\mathrm{BO}]_0 \tag{12.57}$$

Example 12.15 A satellite is operated at an EIRP of 56 dBW with an output BO of 6 dB. The transmitter feeder losses amount to 2 dB, and the antenna gain is 50 dB. Calculate the power output of the TWTA required for full saturated EIRP.

Solution Equation (12.56):

$$\begin{aligned}[P_{\mathrm{TWTA}}] &= [\mathrm{EIRP}]_D - [G_T]_D + [\mathrm{TFL}]_D \\ &= 56 - 50 + 2 \\ &= 8 \text{ dBW}\end{aligned}$$

Equation (12.57):

$$\begin{aligned}[P_{\mathrm{TWTA}}]_S &= 8 + 6 \\ &= \underline{\underline{14 \text{ dBW (or 25 W)}}}\end{aligned}$$

12.9 Effects of Rain

Up to this point, calculations have been made for clear-sky conditions, meaning the absence of weather-related phenomena which might affect the signal strength. In the C band and, more especially, the Ku band, rainfall is the most significant cause of signal fading. Rainfall results in attenuation of radio waves by scattering and by absorption of energy from the wave, as described in Sec. 4.4. Rain attenuation increases with increasing frequency and is worse in the Ku band compared with the C band. Studies have shown (CCIR Report 338-3, 1978) that the rain attenuation for horizontal polarization is considerably greater than for vertical polarization.

Rain attenuation data are usually available in the form of curves or tables showing the fraction of time that a given attenuation is exceeded or, equivalently, the probability that a given attenuation will be exceeded (see Hogg et al., 1975; Lin et al., 1980; Webber et al., 1986). Some yearly average Ku-band values are shown in Table 12.2.

The percentage figures at the head of the first three columns give the percentage of time, averaged over any year, that the attenuation exceeds

TABLE 12.2 Rain Attenuation for Cities and Communities in the Province of Ontario

	Rain attenuation, dB		
Location	1%	0.5%	0.1%
Cat Lake	0.2	0.4	1.4
Fort Severn	0.0	0.1	0.4
Geraldton	0.1	0.2	0.9
Kingston	0.4	0.7	1.9
London	0.3	0.5	1.9
North Bay	0.3	0.4	1.9
Ogoki	0.1	0.2	0.9
Ottawa	0.3	0.5	1.9
Sault Ste. Marie	0.3	0.5	1.8
Sioux Lookout	0.2	0.4	1.3
Sudbury	0.3	0.6	2.0
Thunder Bay	0.2	0.3	1.3
Timmins	0.2	0.3	1.4
Toronto	0.2	0.6	1.8
Windsor	0.3	0.6	2.1

SOURCE: Telesat Canada Design Workbook.

the dB values given in each column. For example, at Thunder Bay, the rain attenuation exceeds, on average throughout the year, 0.2 dB for 1 percent of the time, 0.3 dB for 0.5 percent of the time, and 1.3 dB for 0.1 percent of the time. Alternatively, one could say that for 99 percent of the time, the attenuation will be equal to or less than 0.2 dB; for 99.5 percent of the time, it will be equal to or less than 0.3 dB; and for 99.9 percent of the time, it will be equal to or less than 1.3 dB.

Rain attenuation is accompanied by noise generation, and both the attenuation and the noise adversely affect satellite circuit performance, as described in Secs. 12.9.1 and 12.9.2.

As a result of falling through the atmosphere, raindrops are somewhat flattened in shape, becoming elliptical rather than spherical. When a radio wave with some arbitrary polarization passes through raindrops, the component of electric field in the direction of the major axes of the raindrops will be affected differently from the component along the minor axes. This produces a depolarization of the wave; in effect, the wave becomes elliptically polarized (see Sec. 5.6). This is true for both linear and circular polarizations, and the effect seems to be much worse for circular polarization (Freeman, 1981). Where only a single polarization is involved, the effect is not serious, but where frequency reuse is achieved through the use of orthogonal polarization (as described in Chap. 5), depolarizing devices, which compensate for the rain depolarization, may have to be installed.

Where the earth-station antenna is operated under cover of a radome, the effect of the rain on the radome must be taken into account. Rain

falling on a hemispherical radome forms a water layer of constant thickness. Such a layer introduces losses, both by absorption and by reflection. Results presented by Hogg and Chu (1975) show an attenuation of about 14 dB for a 1-mm-thick water layer. It is desirable, therefore, that earth station antennas be operated without radomes where possible. Without a radome, water will gather on the antenna reflector, but the attenuation produced by this is much less serious than that produced by the wet radome (Hogg and Chu, 1975).

12.9.1 Uplink rain-fade margin

Rainfall results in attenuation of the signal and an increase in noise temperature, degrading the $[C/N_0]$ at the satellite in two ways. The increase in noise, however, is not usually a major factor for the uplink. This is so because the satellite antenna is pointed toward a "hot" earth, and this added to the satellite receiver noise temperature tends to mask any additional noise induced by rain attenuation. What is important is that the uplink carrier power at the satellite must be held within close limits for certain modes of operation, and some form of *uplink power control* is necessary to compensate for rain fades. The power output from the satellite may be monitored by a central control station or in some cases by each earth station, and the power output from any given earth station may be increased if required to compensate for fading. Thus the earth-station HPA must have sufficient reserve power to meet the fade margin requirement.

Some typical rain-fade margins are shown in Table 12.2. As an example, for Ottawa, the rain attenuation exceeds 1.9 dB for 0.1 percent of the time. This means that to meet the specified power requirements at the input to the satellite for 99.9 percent of the time, the earth station must be capable of providing a 1.9-dB margin over the clear-sky conditions.

12.9.2 Downlink rain-fade margin

The results given by Eqs. (12.53) and (12.54) are for clear-sky conditions. Rainfall introduces attenuation by absorption and scattering of signal energy, and the absorptive attenuation introduces noise as discussed in Sec. 12.5.5. Let $[A]$ dB represent the rain attenuation caused by absorption. The corresponding power loss ratio is $A = 10^{[A]/10}$, and substituting this for L in Eq. (12.29) gives the effective noise temperature of the rain as

$$T_{\text{rain}} = T_a\left(1 - \frac{1}{A}\right) \tag{12.58}$$

Here, T_a, which takes the place of T_x in Eq. (12.29), is known as the *apparent absorber temperature.* It is a measured parameter which is a function of many factors including the physical temperature of the rain and the scattering effect of the rain cell on the thermal noise incident upon it (Hogg and Chu, 1975). The value of the apparent absorber temperature lies between 270 and 290 K, with measured values for North America lying close to or just below freezing (273 K). For example, the measured value given by Webber et al. (1986) is 272 K.

The total sky-noise temperature is the clear-sky temperature T_{CS} plus the rain temperature:

$$T_{sky} = T_{CS} + T_{rain} \tag{12.59}$$

Rainfall therefore degrades the received $[C/N_0]$ in two ways: by attenuating the carrier wave and by increasing the sky-noise temperature.

Example 12.16 Under clear-sky conditions, the downlink $[C/N]$ is 20 dB, the effective noise temperature of the receiving system being 400 K. If rain attenuation exceeds 1.9 dB for 0.1 percent of the time, calculate the value below which $[C/N]$ falls for 0.1 percent of the time. Assume $T_a = 280$ K.

Solution 1.9 dB attenuation is equivalent to a 1.55:1 power loss. The equivalent noise temperature of the rain is therefore

$$T_{rain} = 280\,(1 - 1/1.55) = 99.2 \text{ K}$$

The new system noise temperature is 400 + 99.2 = 499.2 K. The decibel increase in noise power is therefore [499.2] − [400] = 0.96 dB. At the same time, the carrier is reduced by 1.9 dB, and therefore, the $[C/N]$ with 1.9-dB rain attenuation drops to 20 − 1.9 − 0.96 = 17.14 dB. This is the value below which $[C/N]$ drops for 0.1 percent of the time.

It is left as an exercise for the student to show that where the rain power attenuation A (not dB) is entirely absorptive, the downlink C/N power ratios (not dBs) are related to the clear-sky value by

$$\left(\frac{N}{C}\right)_{rain} = \left(\frac{N}{C}\right)_{CS}\left(A + (A-1)\frac{T_a}{T_{S,CS}}\right) \tag{12.60}$$

where the subscript CS is used to indicate clear-sky conditions and $T_{S,CS}$ is the system noise temperature under clear-sky conditions. Note that noise-to-carrier ratios, rather than carrier-to-noise ratios are required by Eq. (12.60).

For low frequencies (6/4 GHz) and low rainfall rates (below about 1 mm/h), the rain attenuation is almost entirely absorptive. At higher rainfall rates, scattering becomes significant, especially at the higher frequencies. When scattering and absorption are both significant, the total attenuation must be used to calculate the reduction in carrier power and the absorptive attenuation to calculate the increase in noise temperature.

As discussed in Chap. 9, a minimum value of [C/N] is required for satisfactory reception. In the case of frequency modulation, the minimum value is set by the threshold level of the FM detector, and a *threshold margin* is normally allowed, as shown in Fig. 9.12. Sufficient margin must be allowed so that rain-induced fades do not take the [C/N] below threshold more than a specified percentage of the time, as shown in Example 12.17.

Example 12.17 In an FM satellite system, the clear-sky downlink [C/N] ratio is 17.4 dB and the FM detector threshold is 10 dB, as shown in Fig. 9.12. (*a*) Calculate the threshold margin at the FM detector, assuming the threshold [C/N] is determined solely by the downlink value. (*b*) Given that $T_a = 272$ K and that $T_{S,CS} = 544$ K, calculate the percentage of time the system stays above threshold. The curve of Fig. 12.8 may be used for the downlink, and it may be assumed that the rain attenuation is entirely absorptive.

Solution (*a*) Since it is assumed that the overall [C/N] ratio is equal to the downlink value, the clear-sky input [C/N] to the FM detector is 17.4 dB. The threshold level for the detector is 10 dB, and therefore, the rain-fade margin is 17.4 − 10 = 7.4 dB.

(*b*) The rain attenuation can reduce the [C/N] to the threshold level of 10 dB (i.e., it reduces the margin to zero), which is a (C/N) power ratio of 10:1 or a downlink N/C power ratio of 1/10.

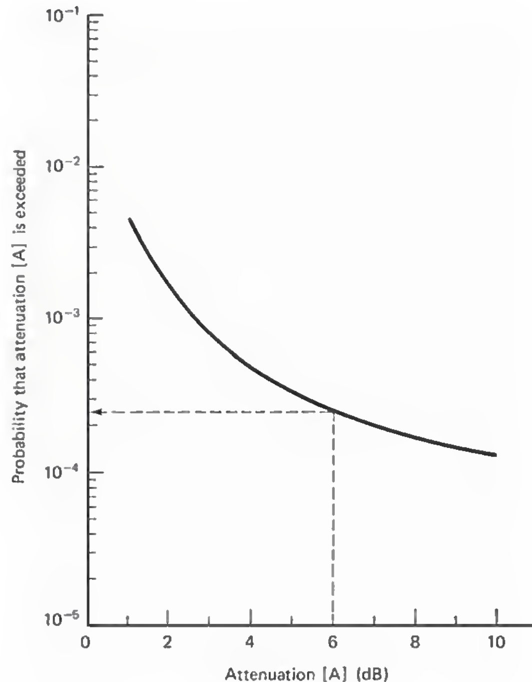


Figure 12.8 Typical rain attenuation curve used in Example 12.17.

For clear-sky conditions, $[C/N]_{CS} = 17.4$ dB, which gives an N/C ratio of 0.0182. Substituting these values in Eq. (12.60) gives

$$0.1 = 0.0182 \times \left(A + \frac{(A - 1) \times 272}{544}\right)$$

Solving this equation for A gives $A = 4$, or approximately 6 dB. From the curve of Fig. 12.8, the probability of exceeding the 6-dB value is 2.5×10^{-4}, and therefore, the availability is $1 - 2.5 \times 10^{-4} = 0.99975$, or 99.975 percent.

For digital signals, the required $[C/N_0]$ ratio is determined by the acceptable BER, which must not be exceeded for more than a specified percentage of the time. Figure 10.17 relates the BER to the $[E_b/N_0]$ ratio, and this in turn is related to the $[C/N_0]$ by Eq. (10.24), as discussed in Sec. 10.6.4.

For the downlink, the user does not have control of the satellite [EIRP], and thus the downlink equivalent of uplink power control, described in Sec. 12.9.1, cannot be used. In order to provide the rain-fade margin needed, the gain of the receiving antenna may be increased by using a larger dish and/or a receiver front end having a lower noise temperature. Both measures increase the receiver $[G/T]$ ratio and thus increase $[C/N_0]$ as shown by Eq. (12.53).

12.10 Combined Uplink and Downlink *C/N* Ratio

The complete satellite circuit includes an uplink and a downlink, as sketched in Fig. 12.9*a*. Noise will be introduced on the uplink at the satellite receiver input. Denoting the noise power per unit bandwidth by P_{NU} and the average carrier at the same point by P_{RU}, the carrier-to-noise ratio on the uplink is $(C/N_0)_U = (P_{RU}/P_{NU})$. It is important to note that power levels, and not decibels, are being used here.

The carrier power at the end of the space link is shown as P_R, which of course is also the received carrier power for the downlink. This is equal to γ times the carrier power input at the satellite, where γ is the system power gain from satellite input to earth-station input, as shown in Fig. 12.9*a*. It includes the satellite transponder and transmit antenna gains, the downlink losses, and the earth-station receive antenna gain and feeder losses.

The noise at the satellite input also appears at the earth station input multiplied by γ, and in addition, the earth station introduces its own noise, denoted by P_{ND}. Thus the end-of-link noise is $\gamma P_{NU} + P_{ND}$.

The C/N_0 ratio for the downlink alone, not counting the γP_{NU} contribution, is P_R/P_{ND}, and the combined C/N_0 ratio at the ground receiver is

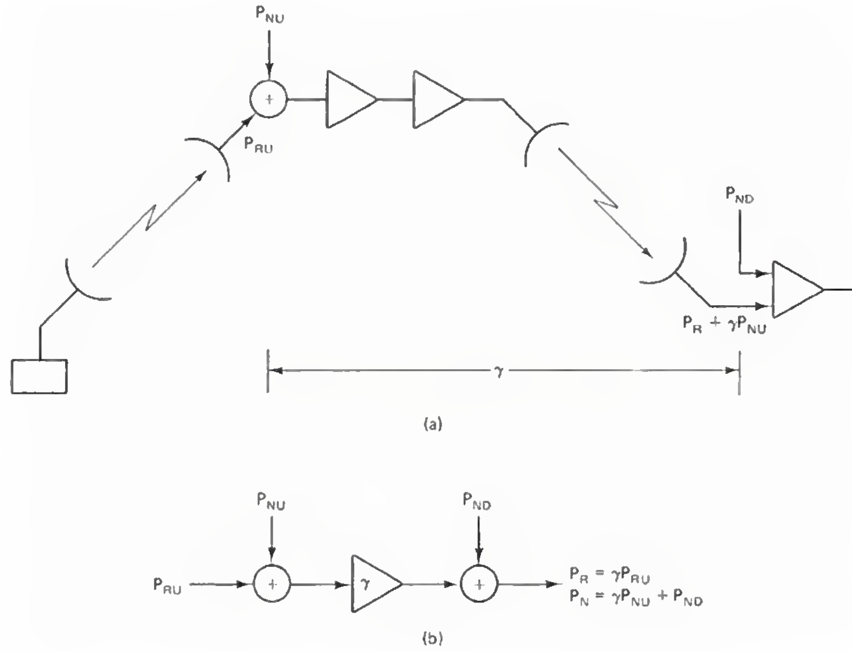


Figure 12.9 (*a*) Combined uplink and downlink; (*b*) power flow diagram for (*a*).

$P_R/(\gamma P_{NU} + P_{ND})$. The power flow diagram is shown in Fig. 12.9*b*. The combined carrier-to-noise ratio can be determined in terms of the individual link values. To show this, it is more convenient to work with the noise-to-carrier ratios rather than the carrier-to-noise ratios, and again, these must be expressed as power ratios, not decibels. Denoting the combined noise-to-carrier ratio value by N_0/C, the uplink value by $(N_0/C)_U$, and the downlink value by $(N_0/C)_D$ then,

$$\begin{aligned}\frac{N_0}{C} &= \frac{P_N}{P_R} \\ &= \frac{\gamma P_{NU} + P_{ND}}{P_R} \\ &= \frac{\gamma P_{NU}}{P_R} + \frac{P_{ND}}{P_R} \\ &= \frac{\gamma P_{NU}}{\gamma P_{RU}} + \frac{P_{NU}}{P_R} \\ &= \left(\frac{N_0}{C}\right)_U + \left(\frac{N_0}{C}\right)_D\end{aligned} \tag{12.61}$$

Equation (12.61) shows that to obtain the combined value of C/N_0, the reciprocals of the individual values must be added to obtain the N_0/C ratio and then the reciprocal of this taken to get C/N_0. Looked at in another way, the reason for this reciprocal of the sum of the reciprocals method is that a single signal power is being transferred through the system, while the various noise powers, which are present are additive. Similar reasoning applies to the carrier-to-noise ratio, C/N.

Example 12.18 For a satellite circuit the individual link carrier-to-noise spectral density ratios are: uplink 100 dBHz; downlink 87 dBHz. Calculate the combined C/N_0 ratio.

Solution

$$\frac{N_0}{C} = 10^{-10} + 10^{-8.7} = 2.095 \times 10^{-9}$$

Therefore,

$$\left[\frac{C}{N_0}\right] = -10 \log(2.095 \times 10^{-9})$$

$$= \underline{\underline{86.79 \text{ dBHz}}}$$

Example 12.18 illustrates the point that when one of the link C/N_0 ratios is much less than the other, the combined C/N_0 ratio is approximately equal to the lower (worst) one. The downlink C/N is usually (but not always) less than the uplink C/N_0, and in many cases it is much less. This is true primarily because of the limited EIRP available from the satellite.

Example 12.19 illustrates how BO is taken into account in the link-budget calculations and how it affects the C/N_0 ratio.

Example 12.19 A multiple carrier satellite circuit operates in the 6/4-GHz band with the following characteristics.

Uplink:

Saturation flux density -67.5 dBW/m^2; input BO 11 dB; satellite G/T -11.6 dBK^{-1}.

Downlink:

Satellite saturation EIRP 26.6 dBW; output BO 6 dB; free-space loss 196.7 dB; earth station G/T 40.7 dBK^{-1}. For this example, the other losses may be ignored. Calculate the carrier-to-noise density ratios for both links and the combined value.

Solution As in the previous examples, the data are best presented in tabular form, and values are shown in decilogs. The minus signs in Eqs. (12.50) and (12.55) are attached to the tabulated numbers:

	Decilog values
Uplink	
Saturation flux density	−67.5
$[A_0]$ at 6 GHz	−37
Input BO	−11
Satellite saturation $[G/T]$	−11.6
$-[k]$	228.6
$[C/N_0]$ from Eq. (12.50)	101.5
Downlink	
Satellite [EIRP]	26.6
Output BO	−6
Free-space loss	−196.7
Earth station $[G/T]$	40.7
$-[k]$	228.6
$[C/N_0]$ from Eq. (12.55)	93.2

Application of Eq. (12.61) provides the combined $[C/N_0]$:

$$\frac{N_0}{C} = 10^{-10.15} + 10^{-9.32} = 5.49 \times 10^{-10}$$

$$\left[\frac{C}{N_0}\right] = -10\log(5.49 \times 10^{-10})$$

$$= \underline{\underline{92.6 \text{ dBHz}}}$$

Again, it is seen from Example 12.19 that the combined C/N_0 value is close to the lowest value, which is the downlink value.

So far, only thermal and antenna noise has been taken into account in calculating the combined value of C/N_0 ratio. Another source of noise to be considered is intermodulation noise, which is discussed in the following section.

12.11 Intermodulation Noise

Intermodulation occurs where multiple carriers pass through any device with nonlinear characteristics. In satellite communications systems, this most commonly occurs in the traveling-wave tube HPA aboard the satellite, as described in Sec. 7.7.3. Both amplitude and phase nonlinearities give rise to intermodulation products.

As shown in Fig. 7.20, third-order intermodulation products fall on neighboring carrier frequencies, where they result in interference. Where a large number of modulated carriers are present, the intermodulation products are not distinguishable separately but instead appear as a type of noise which is termed *intermodulation noise.*

The carrier-to-intermodulation-noise ratio is usually found experimentally, or in some cases it may be determined by computer methods. Once this ratio is known, it can be combined with the carrier-to-thermal-noise ratio by the addition of the reciprocals in the manner described in Sec. 12.10. Denoting the intermodulation term by $(C/N_0)_{\mathrm{IM}}$ and bearing in mind that the reciprocals of the C/N_0 power ratios (and not the corresponding dB values) must be added, Eq. (12.61) is extended to

$$\frac{N_0}{C} = \left(\frac{N_0}{C}\right)_U + \left(\frac{N_0}{C}\right)_D + \left(\frac{N_0}{C}\right)_{IM} \tag{12.62}$$

A similar expression applies for noice-to-carrier (N/C) ratios.

Example 12.20 For a satellite circuit the carrier-to-noise ratios are uplink 23 dB, downlink 20 dB, intermodulation 24 dB. Calculate the overall carrier- to-noise ratio in decibels.

Solution From Eq. (12.62),

$$\frac{N}{C} = 10^{-2.4} + 10^{-2.3} + 10^{-2} = 0.0019$$

$$\left[\frac{C}{N}\right] = -10 \log(0.0019)$$

$$= \underline{\underline{17.2 \text{ dBHz}}}$$

In order to reduce intermodulation noise, the TWT must be operated in a BO condition as described previously. Figure 12.10 shows how the $[C/N_0]_{\mathrm{IM}}$ ratio improves as the input BO is increased for a typical TWT. At the same time, increasing the BO decreases both $[C/N_0]_U$ and $[C/N_0]_D$, as shown by Eqs. (12.50) and (12.55). The result is that there is an optimal point where the overall carrier-to-noise ratio is a maximum. The component $[C/N_0]$ ratios as functions of the TWT input are sketched in Fig. 12.11. The TWT input in dB is $[\Psi]_S - [\mathrm{BO}]_i$, and therefore, Eq. (12.50) plots as a straight line. Equation (12.55) reflects the curvature in the TWT characteristic through the output BO, $[\mathrm{BO}]_0$, which is not linearly related to the input BO, as shown in Fig. 12.7. The intermodulation curve is not easily predictable, and only the general trend is shown. The overall $[C/N_0]$, which is calculated from Eq. (12.62), is also sketched. The optimal operating point is defined by the peak of this curve.

12.12 Inter-Satellite Links

Inter-satellite links (ISLs) are radio frequency or optical links that provide a connection between satellites without the need for intermediate

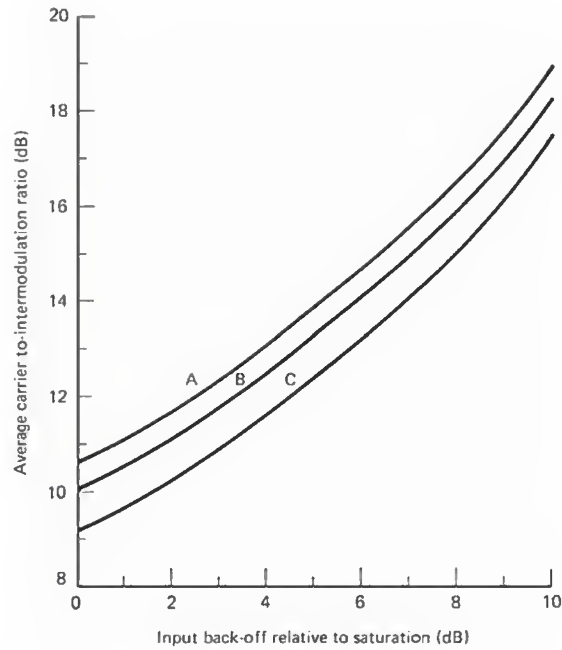


Figure 12.10 Intermodulation in a typical TWT. Curve A, 6 carriers; curve B, 12 carriers; curve C, 500 carriers. (*From CCIR, 1982b. With permission from the International Telecommunications Union.*)

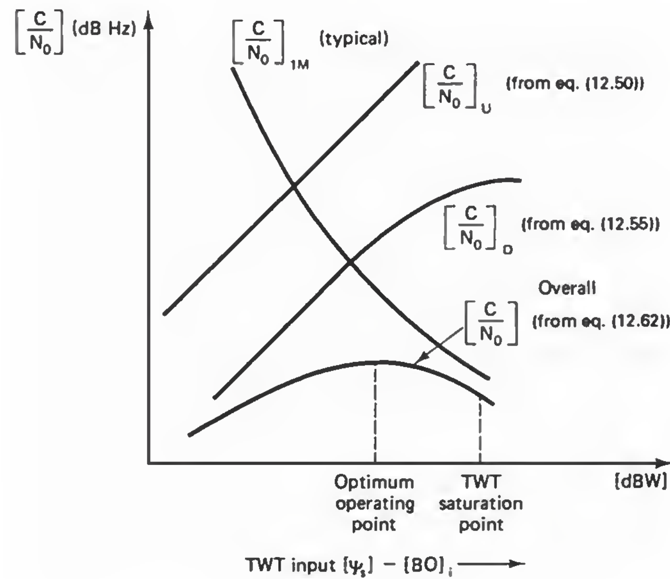


Figure 12.11 Carrier-to-noise density ratios as a function of input back-off.

ground stations. Although many different links are possible, the most useful ones in operation are:

- *low earth orbiting* (LEO) satellites—LEO ↔ LEO
- *geostationary earth orbiting* (GEO) satellites—GEO ↔ GEO
- LEO ↔ GEO

Consider first some of the applications to GEOs. As shown in Chap. 3, the antenna angle of elevation is limited to a minimum of about 5 degrees because of noise induced from the earth. The limit of visibility as set by the minimum angle of elevation is a function of the satellite longitude and earth-station latitude and longitude as shown in Sec. 3.2. Figure 12.12 shows the situation where earth station A is beyond the range of satellite $S2$, a problem that can be overcome by the use of two satellites connected by an ISL. Thus, a long distance link between earth stations A and B can be achieved by this means. A more extreme example is where an intercontinental service may require a number of "hops." For example a Europe-Asia circuit requires three hops (Morgan, 1999): Europe to eastern U.S.A; eastern U.S.A. to western U.S.A; western U.S.A. to Asia; and of course each hop required an uplink and a downlink. By using an ISL only one uplink and one downlink is required. Also, as will be discussed shortly, the ISL frequencies are well outside the standard uplink and downlink bands so that spectrum use is conserved. The cost of the ISL is more than offset by not having to provide the additional earth stations required by the three-hop system.

The distance d for the ISL is easily calculated. From Fig. 12.13:

$$d = 2a_{\text{GSO}} \sin \frac{\Delta\phi}{2} \tag{12.63}$$

where $\Delta\phi$ is the longitudinal separation between satellites $S1$ and $S2$, and a_{GSO} is the radius of the geostationary orbit [see Eq. (3.2)], equal to 42164 km. For example the western limits for the continental United States (CONUS) arc are at 55° and 136° (see Prob. 3.11). Although there are no satellites positioned exactly at these longitudes they can be used

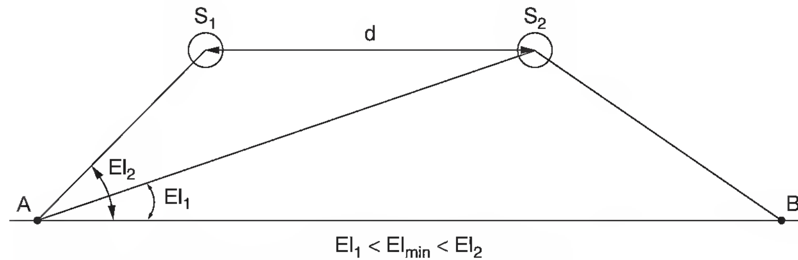


Figure 12.12 Angle of elevation as determined by an ISL.

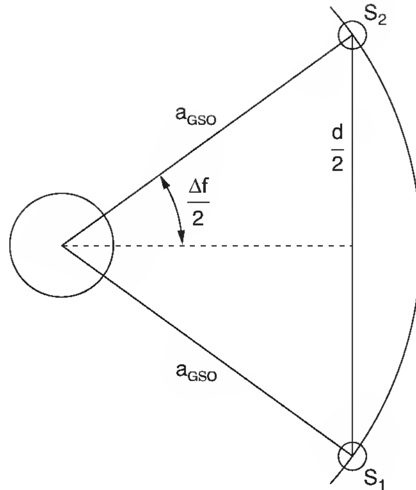


Figure 12.13 Finding the distance d between two GEO satellites.

to get an estimate of distance d for an ISL spanning CONUS.

$$d = 2 \times 42164 \times \sin\frac{(136° - 55°)}{2} = 54767 \text{ km}$$

Although this may seem large, the range from earth station to satellite is in the order of 41000 km, (Prob. 3.11) so distance involved with three uplinks and three downlinks is 246000 km!

GEO satellites are often arranged in clusters at some nominal longitude. For example, there are a number of EchoStar satellites at longitude 119°W. The separation between satellites is typically about 100 km, the corresponding longitudinal separation for this distance being, from Eq. (12.63), approximately 0.136°. Because the satellites are relatively close together they are subject to the same perturbing and drift forces which simplifies positional control. Also, all satellites in the cluster are within the main lobe of the earth-station antenna.

Because LEO satellites are not continuously visible from a given earth location, an intricate network of satellites is required to provide continuous coverage of any region. A typical LEO satellite network will utilize a number of orbits, with equispaced satellites in each orbit. For example, the Iridium system uses 6 orbital planes with 11 equispaced satellites in each plane, for a total of 66 satellites. Communication between two earth stations via the network will appear seamless as message handover occurs between satellites via intersatellite links.

Radio frequency ISLs make use of frequencies that are highly attenuated by the atmosphere, so that interference to and from terrestrial systems using the same frequencies is avoided. Figure 4.2 shows the atmospheric absorption peaks at 22.3 GHz and 60 GHz. Table 12.3 shows the frequency bands in use:

TABLE 12.3 ISL Frequency Bands

Frequency band, GHz	Available bandwidth, MHz	Designation
22.55–23.55	1000	ISL-23
24.45–24.75	300	ISL-24
25.25–27.5	2250	ISL-25
32–33	1000	ISL-32
54.25–58.2	3950	ISL-56
59–64	5000	ISL-60
65–71	6000	ISL-67
116–134	18000	ISL-125
170–182	12000	
185–190	5000	

SOURCE: Morgan, 1999.

Antennas for the ISL are steerable and the beamwidths are sufficiently broad to enable a tracking signal to be acquired to maintain alignment. Table 12.4 gives some values for an ISL used in the Iridium system.

Example 12.21 Calculate the free-space loss for the ISL parameters tabulated in Table 12.4. Given that the system margin for transmission loss is 1.8 dB, calculate the received power.

Solution From Eq. (12.10):

$$[\mathrm{FSL}] = 32.4 + 20 \log r + 20 \log f$$

$$= 32.4 + 20 \log 4400 + 20 \log(23.28 \times 10^3)$$

$$= \underline{\underline{192.6 \text{ dB}}}$$

TABLE 12.4 Iridium ISL

	East-West ISL (without sun)
Frequency, GHz	23.28
Range, km	4400.0
Transmitter	
Power, dBW	5.3
Antenna gain, dB	36.7
Circuit loss, dB	1.8
Pointing loss, dB	1.8
Receiver	
Pointing loss, dB	1.8
Antenna gain, dB	36.7
Noise temperature, K	720.3
Noise bandwidth, dBHz	71

SOURCE: Motorola, 1992.

The [EIRP] is

$$[\text{EIRP}] = [P_T] + [G_T] - [\text{AML}]_T - [\text{TFL}]$$
$$= 5.3 + 36.7 - 1.8 - 1.8$$
$$= 38.4 \text{ dBW}$$

The total losses, including the link margin and the receiver misalignment (pointing) loss are:

[LOSSES] = 192.6 + 1.8 + 1.8 = 196.2 dB. The received power is, from Eq. (12.13)

$$[P_R] = [\text{EIRP}] + [G_R] - [\text{LOSSES}]$$
$$= \underline{\underline{-121.1 \text{ dBW}}}$$

Radio ISLs have the advantage that the technology is mature, so the risk of failure is minimized. However, the bandwidth limits the bit rate that can be carried, and optical systems, with their much higher carrier (optical) frequencies, have much greater bandwidth. Optical ISLs have a definite advantage over rf ISLs for data rates in excess of about 1 Gbps Also, telescope apertures are used which are considerably smaller than their rf counterparts, and generally, optical equipment tends to be smaller and more compact (see Optical Communications and IntersatelliteLinks, undated, at www.wtec.org/loyola/satcom2/03_06.htm-22k-). The optical beamwidth is typically 5 μrad (Maral et al., 2002). Table 12.5 lists properties of some solid state lasers.

The free-space loss given by Eq. (12.9) is repeated here:

$$[\text{FSL}] = 10 \log\left(\frac{4\pi r}{\lambda}\right)^2$$

TABLE 12.5 Solid State Lasers

Type	Wavelength, μm	Power	Beam diameter, mm	Beam divergence
GaAs/GaAlAs	0.78–0.905	1–40 mW Avg		10° × 35°
InGaAsP	1.1–1.6	1–10 mW		10° × 30° − 20° × 40°
Nd: YAG Pulsed	1.064	Up to 600 W Avg	1– 10	0.3–20 mrad
Nd:YAG Diode pumped	1.064	0.5–10 mW	1–2	0.5–2.0 mrad
Nd:YAG (cw)	1.064	0.04–600 W	0.7–8	2–25 mrad

NOTES: Al—aluminum; As—arsenide; Ga—gallium; Nd—neodymium; P—phosphorus; YAG—yttrium-aluminum garnet; cw—continuous wave; μm—micron = 10^{-6} m; mm—millimeter; mrad—milliradian; mW—milliwatt; W—watt.
SOURCE: Extracted from Chen, 1996.

Here, wavelength rather than frequency is used in the equation as this is the quantity usually specified for a laser, and of course r and λ must be in the same units.

The intensity distribution of a laser beam generally follows what is termed a *Gaussian law*, for which the intensity falls off in an exponential manner in a direction transverse to the direction of propagation. The *beam radius* is where the transverse electric field component drops to 1/e of its maximum value, where e ≈ 2.718. The diameter of the beam (twice the radius) gives the total beamwidth. The on-axis gain (similar to the antenna gain defined in Sec. 6.6 is given by (Maral et al., 2002)

$$G_T = \frac{32}{\theta_T^2} \tag{12.64}$$

where θ_T is the total beamwidth.

On the receive side, the telescope aperture gain is given by:

$$G_R = \left(\frac{\pi D}{\lambda}\right)^2 \tag{12.65}$$

where D is the effective diameter of the receiving aperture.

The optical receiver will receive some amount of optical power P_R. The energy in a photon is hc/λ, where h is Plank's constant (6.6256×10^{-34} J-s) and c is the speed of light in vacuum (approximately 3×10^8 m/s). For a received power P_R the number of photons received per second is therefore $P_R\lambda/hc$. The detection process consists of photons imparting sufficient energy to valence band electrons to raise these to the conduction band. The *quantum efficiency* of a photo-diode is the ratio (average number of conduction electrons generated)/(average number of photons received). Denoting the quantum efficiency by η, the average number of electrons released is $\eta P_R\lambda/hc$ and the photo current is:

$$I_{ph} = \frac{q\eta P_R\lambda}{hc} \tag{12.66}$$

where q is the electron charge. The *responsivity* of a photodiode is defined as the ratio of photo current to incident power. Denoting responsivity by R_0 and evaluating the constants in Eq. (12.66) gives

$$R_0 = \frac{\eta\lambda}{1.24} \quad \text{with } \lambda \text{ in } \mu\text{m} \tag{12.67}$$

The energy band gap of the semiconductor material used for the photodiode determines the wavelengths that it can respond to. The requirement in general is that the bandgap energy must be less than the photon

energy, or $E_G < hc/\lambda$. On this basis it turns out that silicon is useful for wavelengths shorter that 1 μm. Germanium is useful at 1.3 μm and InSb and InAs at 1.55 μm.

The total current flowing in a photodiode consists of the actual current generated by the photons, plus what is termed the *dark current*. This is the current that flows even when no signal is present. Ideally it should be zero, but in practice it is of the order of a few nanoamperes, and it can contribute to the noise. Denoting the dark current by I_d the total current is

$$I = I_{ph} + I_d \tag{12.68}$$

This current is accompanied by *shot noise* (a name that is a hangover from vacuum tube days), the mean square spectral density of which is $2qI$ in A^2/Hz with I in amperes. This would be for a diode without any internal amplification such as a PIN diode. An *avalanche photo diode* (APD) multiplies the signal current by a factor M and at the same time generates excess noise, represented by an *excess noise factor, F*, so that the mean square spectral density is $(2qI)MF$. F increases with increase in M (see Jones, 1988, p. 240). Typical values are of the order $M = 100$, $F = 12$. For a PIN diode, $M = 1$ and $F = 1$.

The analysis to follow is based on that given in Jones (1988). The equivalent circuit for the input stage of an optical receiver is shown in Fig. 12.14, where R is the parallel combination of diode resistance, input load resistance and preamplifier input resistance, and C is the parallel combination of diode capacitance, preamplifier input capacitance and stray circuit capacitance. Resistance R will generate thermal noise current, the spectral density of which is $4kT/R$ (k is Boltzmann's constant and T is the temperature, which may be taken as room temperature). The preamplifier transistor will also generate shot noise, which when

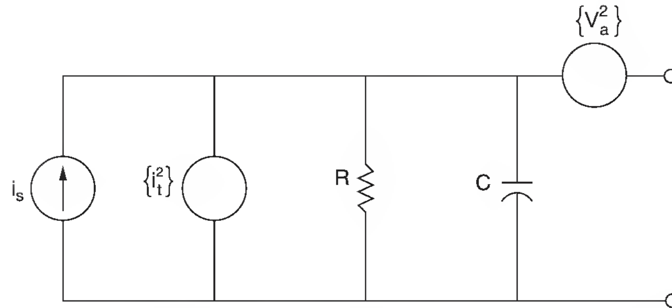


Figure 12.14 The equivalent input circuit for an optical receiver. $\{i_t^2\}$ is the mean square spectral density for the current noise source, and $\{v_a^2\}$ the mean square spectral density for the voltage noise source. i_s is the signal current.

referred to the input has a spectral density $2qI_{in}$ in A^2/Hz, where I_{in} is the transistor gate (JFET) or base (BJT) current in amperes. Using curly brackets {.} for the spectral density values (see Jones, 1988), the total noise current spectral density at the input is

$$\{i_t^2\} = 2qIM^2F + \frac{4kT}{R} + 2qI_{tr} \ \text{A}^2/\text{Hz} \tag{12.69}$$

The pre-amplifier also has a noise voltage component shown as v_a resulting from the shot noise in the drain or collector current. The mean square noise voltage spectral density is $2qI_{tr}/g_m^2$ V^2/Hz where I_{tr} is the transistor drain (JFET) or collector (BJT) current in amperes and g_m is the device transconductance in siemens. The total noise voltage spectral density at the input is therefore

$$\{v_n^2\} = \{i_t^2\}|Z_L|^2 + \frac{2qI_{tr}}{g_m^2} \ \text{V}^2/\text{Hz} \tag{12.70}$$

where Z_L is the impedance of R and C in parallel.

The average signal voltage is $MI_{ph}R$. Increasing R should result in an increase in signal to noise ratio since the signal voltage is proportional to R and the R component of noise current density is inversely proportional to R. However, the bandwidth of the RC input network is inversely proportional to R and therefore sets a limit to how large R can be. An equalizing filter, which has a transfer function, given by $H_{eq}(f) = 1 + j2\pi f\text{RC}$ can be included in the overall transfer function, which compensates for the input impedance frequency response over the signal bandwidth. The effective spectral density for the mean square noise voltage at the input is then $\{v_n^2\}|H(f)|^2$. The mean square noise voltage V_n^2 at the input is obtained by integrating this expression over the signal bandwidth. Only the result will be given here:

$$V_n^2 = \left[\left(2qIM^2F + \frac{4kT}{R} + 2qI_{tr}\right)R^2 + \frac{2qI_{tr}}{g_m^2}\left(1 + \frac{(2\pi\text{RCB})^2}{3}\right)\right]B \tag{12.71}$$

With the average signal voltage given by $V_s = MI_{ph}R$ the signal to noise ratio is

$$\frac{S}{N} = \frac{V_s^2}{V_n^2}$$

$$= \frac{(MI_{ph}R)^2}{\left[\left(2qIM^2F + \frac{4kT}{R} + 2qI_{in}\right)R^2 + \frac{2qI_{tr}}{g_m^2}\left(1 + \frac{(2\pi RCB)^2}{3}\right)\right]B} \tag{12.72}$$

This can be simplified on dividing through by M^2R^2 to get:

$$\frac{S}{N}=\frac{I_{ph}^2}{\left[2qIF+\dfrac{4kT}{M^2R}+\dfrac{2qI_{in}}{M^2}+\dfrac{2qI_{tr}}{M^2g_m^2}\left(\dfrac{1}{R^2}+\dfrac{(2\pi CB)^2}{3}\right)\right]B} \tag{12.73}$$

Example 12.22 An optical receiver utilizes an APD for which $R_0 = 0.65$ A/W, $M = 100$, $F = 4$, $I_{in} = 0$, gm = 3000 mS and $I_{tr} = 0.15$ mA. The dark current may be neglected. The input load consists of a 600 Ω resistor in parallel with a 10pF capacitance. The signal bandwidth is 25 MHz and equalization is employed. Calculate the resultant signal-to-noise ratio for an input signal power of 1 μW.

Solution The photocurrent is $I_{ph} = 0.65 \times 10^{-6} = 0.65$ μA. The individual terms in the denominator are, with $I_d = 0$ and $I_{in} = 0$:

$$2qIF = 6.408 \times 10^{-25}\ \mathrm{A^2/Hz}$$

$$\frac{4kT}{M^2R} = 0.027 \times 10^{-25}\ \mathrm{A^2/Hz}$$

$$\frac{2qI_{tr}}{M^2g_m^2}\left(\frac{1}{R^2}+\frac{(2\pi \mathrm{CB})^2}{3}\right) = 0.011 \times 10^{-25}\ \mathrm{A^2/Hz}$$

$$\frac{S}{N}=\frac{(0.65 \times 10^{-6})^2 \times 10^{25}}{(6.408 + 0.027 + 0.011) \times 25 \times 10^6}$$

$$= 2.622 \times 10^4$$

In decibels this is 44.2 dB

12.13 Problems and Exercises

Note: In problems where room temperature is required, assume a value of 290 K. In calculations involving antenna gain, an efficiency factor of 0.55 may be assumed.

12.1. Give the decibel equivalents for the following quantities: (*a*) a power ratio of 30:1; (*b*) a power of 230 W; (*c*) a bandwidth of 36 MHz; (*d*) a frequency ratio of 2 MHz/3 kHz; (*e*) a temperature of 200 K.

12.2. (*a*) Explain what is meant by EIRP. (*b*) A transmitter feeds a power of 10 W into an antenna which has a gain of 46 dB. Calculate the EIRP in (*i*) watts; (*ii*) dBW.

12.3. Calculate the gain of a 3-m parabolic reflector antenna at a frequency of (*a*) 6 GHz; (*b*) 14 GHz.

12.4. Calculate the gain in decibels and the effective area of a 30-m parabolic antenna at a frequency of 4 GHz.

12.5. An antenna has a gain of 46 dB at 12 GHz. Calculate its effective area.

12.6. Calculate the effective area of a 10-ft parabolic reflector antenna at a frequency of (*a*) 4 GHz; (*b*) 12 GHz.

12.7. The EIRP from a satellite is 49.4 dBW. Calculate (*a*) the power density at a ground station for which the range is 40,000 km and (*b*) the power delivered to a matched load at the ground station receiver if the antenna gain is 50 dB. The downlink frequency is 4 GHz.

12.8. Calculate the free-space loss as a power ratio and in decibels for transmission at frequencies of (*a*) 4 GHz, (*b*) 6 GHz, (*c*) 12 GHz, and (*d*) 14 GHz; the range being 42,000 km.

12.9. Repeat the calculation in Prob. 12.7*b* allowing for a fading margin of 1.0 dB and receiver feeder losses of 0.5 dB.

12.10. Explain what is meant by (*a*) *antenna noise temperature,* (*b*) *amplifier noise temperature*, and (*c*) *system noise temperature* referred to input. A system operates with an antenna noise temperature of 40 K and an input amplifier noise temperature of 120 K. Calculate the available noise power density of the system referred to the amplifier input.

12.11. Two amplifiers are connected in cascade, each having a gain of 10 dB and a noise temperature of 200 K. Calculate (*a*) the overall gain and (*b*) the effective noise temperature referred to input.

12.12. Explain what is meant by *noise factor*. For what source temperature is noise factor defined?

12.13. The noise factor of an amplifier is 7:1. Calculate (*a*) the noise figure and (*b*) the equivalent noise temperature.

12.14. An attenuator has an attenuation of 6 dB. Calculate (*a*) its noise figure and (*b*) its equivalent noise temperature referred to input.

12.15. An amplifier having a noise temperature of 200 K has a 4-dB attenuator connected at its input. Calculate the effective noise temperature referred to the attenuator input.

12.16. A receiving system consists of an antenna having a noise temperature of 60 K, feeding directly into a LNA. The amplifier has a noise temperature of 120 K and a gain of 45 dB. The coaxial feeder between the LNA and the main receiver has a loss of 2 dB, and the main receiver has a noise figure of 9 dB. Calculate the system noise temperature referred to input.

12.17. Explain why the LNA of a receiving system is placed at the antenna end of the feeder cable.

12.18. An antenna having a noise temperature of 35 K is connected through a feeder having 0.5-dB loss to an LNA. The LNA has a noise temperature of 90 K. Calculate the system noise temperature referred to (*a*) the feeder input and (*b*) the LNA input.

12.19. Explain what is meant by *carrier-to-noise ratio*. At the input to a receiver the received carrier power is 400 pW and the system noise temperature is 450 K. Calculate the carrier-to-noise density ratio in dBHz. Given that the bandwidth is 36 MHz, calculate the carrier-to-noise ratio in decibels.

12.20. Explain what is meant by the G/T ratio of a satellite receiving system. A satellite receiving system employs a 5-m parabolic antenna operating at 12 GHz. The antenna noise temperature is 100 K, and the receiver front-end noise temperature is 120 K. Calculate $[G/T]$.

12.21. In a satellite link the propagation loss is 200 dB. Margins and other losses account for another 3 dB. The receiver $[G/T]$ is 11 dB, and the [EIRP] is 45 dBW. Calculate the received $[C/N]$ for a system bandwidth of 36 MHz.

12.22. A carrier-to-noise density ratio of 90 dBHz is required at a receiver having a $[G/T]$ ratio of 12 dB. Given that total losses in the link amount to 196 dB, calculate the [EIRP] required.

12.23. Explain what is meant by *saturation flux density*. The power received by a 1.8-m parabolic antenna at 14 GHz is 250 pW. Calculate the power flux density (*a*) in W/m^2 and (*b*) in dBW/m^2 at the antenna.

12.24. An earth station radiates an [EIRP] of 54 dBW at a frequency of 6 GHz. Assuming that total losses amount to 200 dB, calculate the power flux density at the satellite receiver.

12.25. A satellite transponder requires a saturation flux density of -110 dBW/m^2, operating at a frequency of 14 GHz. Calculate the earth station [EIRP] required if total losses amount to 200 dB.

12.26. Explain what is meant by input BO. An earth station is required to operate at an [EIRP] of 44 dBW in order to produce saturation of the satellite transponder. If the transponder has to be operated in a 10 dB input BO mode, calculate the new value of [EIRP] required.

12.27. Determine the carrier-to-noise density ratio at the satellite input for an uplink, which has the following parameters: operating frequency 6 GHz, saturation flux density -95 dBW/m^2, input BO 11 dB, satellite $[G/T]$ -7 dBK^{-1}, [RFL] 0.5 dB. (Tabulate the link budget values as shown in the text).

12.28. For an uplink the required [C/N] ratio is 20 dB. The operating frequency is 30 GHz, and the bandwidth is 72 MHz. The satellite [G/T] is 14.5 dBK^{-1}. Assuming operation with 11 dB input BO, calculate the saturation flux density. [RFL] are 1 dB.

12.29. For the uplink in Prob. 12.28, the total losses amount to 218 dB. Calculate the earth station [EIRP] required.

12.30. An earth station radiates an [EIRP] of 54 dBW at 14 GHz from a 10-m parabolic antenna. The transmit feeder losses between the HPA and the antenna are 2.5 dB. Calculate the output of the HPA.

12.31. The following parameters apply to a satellite downlink: saturation [EIRP] 22.5 dBW, free-space loss 195 dB, other losses and margins 1.5 dB, earth station [G/T] 37.5 dB/K. Calculate the [C/N_0] at the earth station. Assuming an output BO of 6 dB is applied, what is the new value of [C/N_0]?

12.32. The output from a satellite TWTA is 10 W. This is fed to a 1.2-m parabolic antenna operating at 12 GHz, the feeder loss being 2 dB. Calculate the [EIRP].

12.33. The [C/N] values for a satellite circuit are uplink 25 dB, downlink 15 dB. Calculate the overall [C/N] value.

12.34. The required [C/N] value at the ground station receiver is 22 dB and the downlink [C/N] is 24 dB. What is the minimum value of [C/N] that the uplink can have in order that the overall value can be achieved?

12.35. A satellite circuit has the following parameters:

	Uplink, decilogs	Downlink, decilogs
[EIRP]	54	34
[G/T]	0	17
[FSL]	200	198
[RFL]	2	2
[AA]	0.5	0.5
[AML]	0.5	0.5

Calculate the overall [C/N_0] value.

12.36. Explain how intermodulation noise originates in a satellite link, and describe how it may be reduced. In a satellite circuit the carrier-to-noise ratios are uplink 25 dB; downlink 20 dB; intermodulation 13 dB. Calculate the overall carrier-to-noise ratio.

12.37. For the satellite circuit of Prob. 12.36, input BO is introduced that reduces the carrier-to-noise ratios by the following amounts: uplink 7 dB, downlink 2 dB, and improves the intermodulation by 11 dB. Calculate the new value of overall carrier-to-noise ratio.

12.38. As a follow-up to Example 12.21, calculate $[E_b/N_0]$, given that the coded bit rate is 25 Mbps.

12.39. Rework Example 12.22 to find the output [S/N] for a PIN diode, all other values remaining unchanged.

References

Andrew Antenna Co., Ltd. 1985. *1.8-Meter 12-GHz Receive-only Earth Station Antenna. Bulletin 1206A*. Whitby, Ontario, Canada.

CCIR Report 338-3. 1978. "Propagation Data Required for Line of Sight Radio Relay Systems." *14th Plenary Assembly*, Vol. V, Kyoto.

Chen, C. L. 1996. *Elements of Optoelectronics and Fiber Optics*. Irwin, a Times Mirror Higher Education Group, Inc., Chicago.

Freeman, R. L. 1981. *Telecommunications Systems Engineering*. Wiley, New York.

Hogg, D. C., and T. Chu. 1975. "The Role of Rain in Satellite Communications." *Proc. IEEE*, Vol. 63, No. 9, pp. 1308–1331.

Jones, W. B. Jr. 1988. *Introduction to Optical Fiber Communication Systems*. Holt, Rinehart and Winston, lnc., TX.

Lin, S. H., H. J. Bergmann, and M. V. Pursley. 1980. "Rain Attenuation on Earth-Satellite Paths: Summary of 10-Year Experiments and Studies." *Bell Syst. Tech. J.*, Vol. 59, No. 2, February, pp. 183–228.

Maral, G., and M. Bousquet. 2002. *Satellite Communications Systems*. Wiley, New York.

Morgan, W. L. 1999. "Intersatellite Links." *Space Business International*, Quarter 1, pp. 9–12.

Motorola Satellite Communications Inc., 1992. Minor amendment to application before the FCC to construct and operate a low earth orbit satellite system in the RDSS uplink band File No.9-DSS-P91 (87) CSS-91-010.

Webber, R. V., J. I. Strickland, and J. J. Schlesak. 1986. "Statistics of Attenuation by Rain of 13-GHz Signals on Earth-Space Paths in Canada." *CRC Report 1400*, Communications Research Centre, Ottawa, April.

Chapter

13

Interference

13.1 Introduction

With many telecommunications services using radio transmissions, interference between services can arise in a number of ways. Figure 13.1 shows in a rather general way the possible interference paths between services. It will be seen in Fig. 13.1 that the terms *earth station* and *terrestrial station* are used, and the distinction must be carefully noted. Earth stations are specifically associated with satellite circuits, and terrestrial stations are specifically associated with ground-based microwave line-of-sight circuits. The possible modes of interference shown in Fig. 13.1 are classified by the *International Telecommunications Union* (ITU, 1985) as follows:

A_1: terrestrial station transmissions, possibly causing interference to reception by an earth station

A_2: earth station transmissions, possibly causing interference to reception by a terrestrial station

B_1: space station transmission of one space system, possibly causing interference to reception by an earth station of another space system

B_2: earth station transmissions of one space system, possibly causing interference to reception by a space station of another space system

C_1: space station transmission, possibly causing interference to reception by a terrestrial station

C_2: terrestrial station transmission, possibly causing interference to reception by a space station

E: space station transmission of one space system, possibly causing interference to reception by a space station of another space system

F: earth station transmission of one space system, possibly causing interference to reception by an earth station of another space system

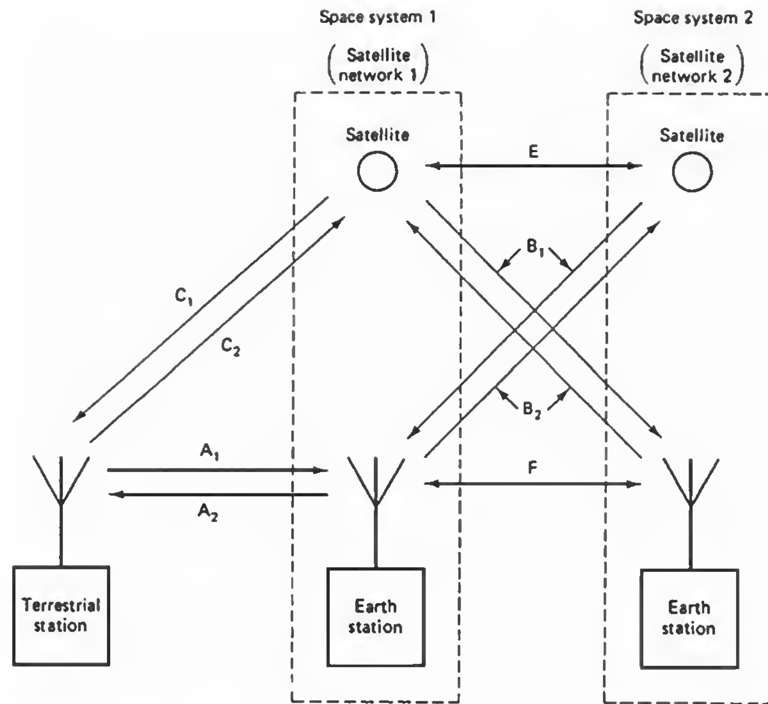


Figure 13.1 Possible interference modes between satellite circuits and a terrestrial station. (*Courtesy of CCIR Radio Regulations.*)

A_1, A_2, C_1, and C_2 are possible modes of interference between space and terrestrial services. B_1 and B_2 are possible modes of interference between stations of different space systems using separate uplink and downlink frequency bands, and E and F are extensions to B_1 and B_2 where bidirectional frequency bands are used.

The Radio Regulations (ITU, 1986) specify maximum limits on radiated powers (more strictly, on the distribution of energy spectral density) in an attempt to reduce the potential interference to acceptable levels in most situations. However, interference may still occur in certain cases, and what is termed *coordination* between the telecommunications administrations that are affected is then required. Coordination may require both administrations to change or adjust some of the technical parameters of their systems.

For geostationary satellites, interference modes B_1 and B_2 set a lower limit to the orbital spacing between satellites. To increase the capacity of the geostationary orbit, the *Federal Communications Commission* (FCC) in the United States has in recent years authorized a reduction in orbital spacing from 4° to 2° in the 6/4-GHz band. Some

of the effects that this has on the B_1 and B_2 levels of interference are examined later in the chapter. It may be noted, however, that although the larger authorized operators will in general be able to meet the costs of technical improvements needed to offset the increased interference resulting from reduced orbital spacing, the same cannot be said for individually owned *television receive-only* (TVRO) installations (the "home satellite dish"), and these users will have no recourse to regulatory control (Chouinard, 1984).

Interference with individually owned TVRO receivers also may occur from terrestrial station transmissions in the 6/4-GHz band. Although this may be thought of as an A_1 mode of interference, the fact that these home stations are considered by many broadcasting companies to be "pirates" means that regulatory controls to reduce interference are not applicable. Some steps that can be taken to reduce this form of interference are described in a publication by the Microwave Filter Company (1984).

It has been mentioned that the Radio Regulations place limits on the energy spectral density which may be emitted by an earth station. Energy dispersal is one technique employed to redistribute the transmitted energy more evenly over the transmitted bandwidth. This principle is described in more detail later in this chapter.

Intermodulation interference, briefly mentioned in Sec. 7.7.3, is a type of interference which can occur between two or more carriers using a common transponder in a satellite or a common high-power amplifier in an earth station. For all practical purposes, this type of interference can be treated as noise, as described in Sec. 12.11.

13.2 Interference between Satellite Circuits (B_1 and B_2 Modes)

A satellite circuit may suffer the B_1 and B_2 modes of interference shown in Fig. 13.1 from a number of neighboring satellite circuits, the resultant effect being termed *aggregate interference*. Because of the difficulties of taking into account the range of variations expected in any practical aggregate, studies of aggregate interference have been quite limited, with most of the study effort going into what is termed *single-entry interference studies* (see Sharp, 1984*a*). As the name suggests, single-entry interference refers to the interference produced by a single interfering circuit on a neighboring circuit.

Interference may be considered as a form of noise, and as with noise, system performance is determined by the ratio of wanted to interfering powers, in this case the wanted carrier to the interfering carrier power or C/I ratio. The single most important factor controlling interference is the radiation pattern of the earth-station antenna. Comparatively

large-diameter reflectors can be used with earth-station antennas, and hence narrow beamwidths can be achieved. For example, a 10-m antenna at 14 GHz has a -3-dB beamwidth of about 0.15°. This is very much narrower than the 2° to 4° orbital spacing allocated to satellites. To relate the *C*/*I* ratio to the antenna radiation pattern, it is necessary first to define the geometry involved.

Figure 13.2 shows the angles subtended by two satellites in geostationary orbit. The *orbital separation* is defined as the angle α subtended at the center of the earth, known as the *geocentric angle*. However, from an earth station at point P the satellites would appear to subtend an angle β. Angle β is referred to as the *topocentric angle*. In all practical situations relating to satellite interference, the topocentric and geocentric angles may be assumed equal, and in fact, making this assumption leads to an overestimate of the interference (Sharp, 1983).

Consider now S_1 as the wanted satellite and S_2 as the interfering satellite. An antenna at P will have its main beam directed at S_1 and an off-axis component at angle θ directed at S_2. Angle θ is the same as the topocentric angle, which as already shown may be assumed equal to the geocentric or orbital spacing angle. Therefore, when calculating the antenna sidelobe pattern, the orbital spacing angle may be used, as described in Sec. 13.2.4. Orbital spacing angles range from 2° to 4° in 0.5° intervals in the C band.

In Fig. 13.3 the satellite circuit being interfered with is that from earth station A via satellite S_1 to receiving station B. The B_1 mode of interference can occur from satellite S_2 into earth station B, and the B_2 mode of interference can occur from earth station C into satellite S_1. The total single-entry interference is the combined effect of these two modes. Because the satellites cannot carry very large antenna reflectors, the beamwidth is relatively wide, even for the so-called spot beams. For example, a 3.5-m antenna at 12 GHz has a beamwidth of about 0.5°,

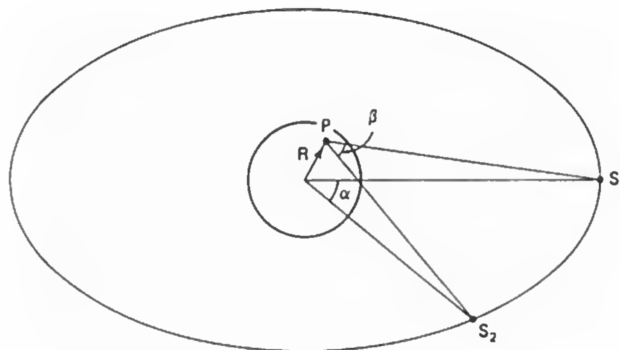


Figure 13.2 Geocentric angle α and the topocentric angle β.

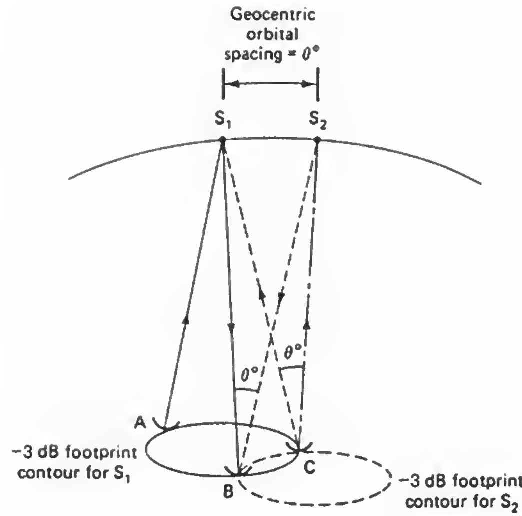


Figure 13.3 Orbital spacing angle.

and the equatorial arc subtended by this angle is about 314 km. In interference calculations, therefore, the earth stations will be assumed to be situated on the −3-dB contours of the satellite footprints, in which case the satellite antennas do not provide any gain discrimination between the wanted and the interfering carriers on either transmit or receive.

13.2.1 Downlink

Equation (12.13) may be used to calculate the wanted and interfering downlink carrier powers received by an earth station. The carrier power $[C]$ in dBW received at station B is

$$[C] = [\text{EIRP}]_1 - 3 + [G_B] - [\text{FSL}] \tag{13.1}$$

Here, $[\text{EIRP}]_1$ is the equivalent isotropic radiated power in dBW from satellite 1, the −3 dB accounts for the −3-dB contour of the satellite transmit antenna, G_B is the boresight (on-axis) receiving antenna gain at B, and [FSL] is the free-space loss in decibels. A similar equation may be used for the interfering carrier $[I]$, except an additional term $[Y]_D$ dB, allowing for polarization discrimination, must be included. Also, the receiving antenna gain at B is determined by the off-axis angle θ, giving

$$[I] = [\text{EIRP}]_2 - 3 + [G_B(\theta)] - [\text{FSL}] - [Y]_D \tag{13.2}$$

It is assumed that the free-space loss is the same for both paths.

These two equations may be combined to give

$$[C] - [I] = [\text{EIRP}]_1 - [\text{EIRP}]_2 + [G_B] - [G_B(\theta)] + [Y]_D$$

or

$$\left[\frac{C}{I}\right]_D = \Delta[E] + [G_B] - [G_B(\theta)] + [Y]_D \tag{13.3}$$

The subscript D is used to denote downlink, and $\Delta[E]$ is the difference in dB between the [EIRP]s of the two satellites.

Example 13.1 The desired carrier [EIRP] from a satellite is 34 dBW, and the ground station receiving antenna gain is 44 dB in the desired direction and 24.47 dB toward the interfering satellite. The interfering satellite also radiates an [EIRP] of 34 dBW. The polarization discrimination is 4 dB. Determine the carrier-to-interference ratio at the ground receiving antenna.

Solution From Eq. (13.3)

$$\left[\frac{C}{I}\right]_D = (34 - 34) + 44 - 24.47 + 4$$

$$= \underline{\underline{23.3 \text{ dB}}}$$

13.2.2 Uplink

A result similar to Eq. (13.3) can be derived for the uplink. In this situation, however, it is desirable to work with the radiated powers and the antenna transmit gains rather than the EIRPs of the two earth stations. Equation (12.3) may be used to substitute power and gain for EIRP. Also, for the uplink, G_B and $G_B(\theta)$ are replaced by the satellite receive antenna gains, both of which are assumed to be given by the -3-dB contour. Denoting by $\Delta[P]$ the difference in dB between wanted and interfering transmit powers, $[G_A]$ the boresight transmit antenna gain at A, and $[G_C(\theta)]$ the off-axis transmit gain at C, it is left as an exercise for the reader to show that Eq. (13.3) is modified to

$$\left[\frac{C}{I}\right]_U = \Delta[P] + [G_A] - [G_C(\theta)] + [Y]_U \tag{13.4}$$

Example 13.2 Station A transmits at 24 dBW with an antenna gain of 54 dB, and station C transmits at 30 dBW. The off-axis gain in the S_1 direction is 24.47 dB, and the polarization discrimination is 4 dB. Calculate the $[C/I]$ ratio on the uplink.

Solution Equation (13.4) gives

$$\left[\frac{C}{I}\right]_D = (24 - 30) + 54 - 24.47 + 4$$

$$= \underline{\underline{27.53 \text{ dB}}}$$

13.2.3 Combined [*C/I*] due to interference on both uplink and downlink

Interference may be considered as a form of noise, and assuming that the interference sources are statistically independent, the interference powers may be added directly to give the total interference at receiver *B*. The uplink and the downlink ratios are combined in exactly the same manner described in Sec. 12.10 for noise, resulting in

$$\left(\frac{I}{C}\right)_{\text{ant}} = \left(\frac{I}{C}\right)_U + \left(\frac{I}{C}\right)_D \tag{13.5}$$

Here, power ratios must be used, not decibels, and the subscript "ant" denotes the combined ratio at the output of station *B* receiving antenna.

Example 13.3 Using the uplink and downlink values of [*C/I*] determined in Examples 13.1 and 13.2, find the overall ratio $[C/I]_{\text{ant}}$.

Solution For the uplink, $[C/I] = 27.53$ dB gives $(I/C)_U = 0.001766$, and for the downlink, $[C/I] = 23.53$ dB gives $(I/C)_D = 0.004436$. Combining these according to Eq. (13.5) gives

$$\left(\frac{I}{C}\right)_{\text{ant}} = 0.001766 + 0.004436$$
$$= 0.006202$$

Hence

$$\left[\frac{C}{I}\right]_{\text{ant}} = -10\log 0.006202$$
$$= \underline{\underline{20.07 \text{ dB}}}$$

13.2.4 Antenna gain function

The antenna radiation pattern can be divided into three regions: the mainlobe region, the sidelobe region, and the transition region between the two. For interference calculations, the fine detail of the antenna pattern is not required, and an envelope curve is used instead.

Figure 13.4 shows a sketch of the envelope pattern used by the FCC. The width of the mainlobe and transition region depend on the ratio of the antenna diameter to the operating wavelength, and Fig. 13.4 is intended to show only the general shape. The sidelobe gain function in decibels is defined for different ranges of θ. Specifying θ in degrees, the

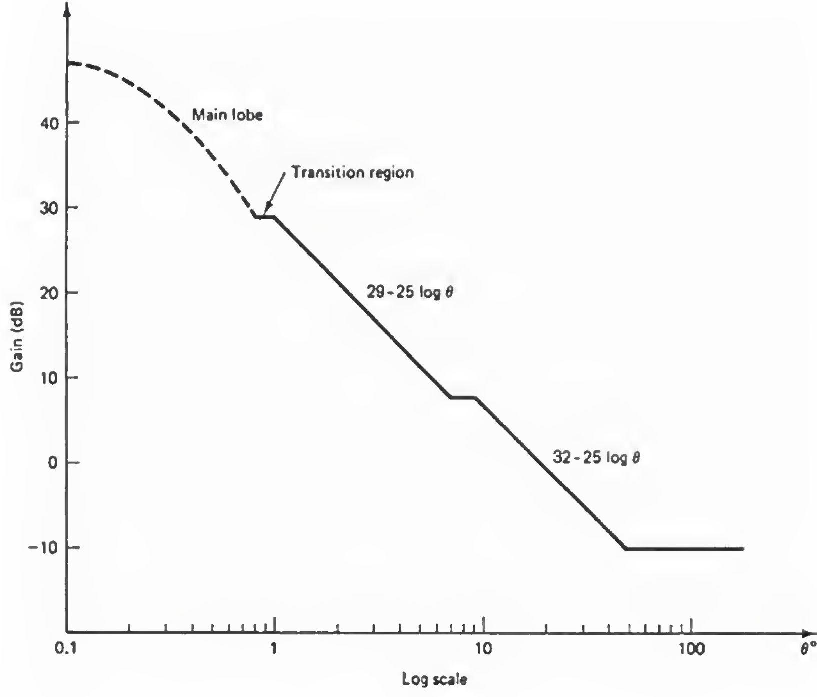


Figure 13.4 Earth-station antenna gain pattern used in FCC/OST R83-2, revised Nov. 30, 1984. (*Courtesy of Sharp, 1984b.*)

sidelobe gain function can be written as follows:

$$[G(\theta)] = \begin{cases} 29 - 25\log\theta & 1 \le \theta \le 7 \\ +8 & 7 < \theta \le 9.2 \\ 32 - 25\log\theta & 9.2 < \theta \le 48 \\ -10 & 48 < \theta \le 180 \end{cases} \tag{13.6}$$

For the range of satellite orbital spacings presently in use, it is this sidelobe gain function that determines the interference levels.

Example 13.4 Determine the degradation in the downlink $[C/I]$ ratio when satellite orbital spacing is reduced from 4° to 2°, all other factors remaining unchanged. FCC antenna characteristics may be assumed.

Solution The decibel increase in interference will be

$$(29 - 25 \log 2) - (29 - 25 \log 4) = 25 \log 2 = \underline{\underline{7.5 \text{ dB}}}$$

The $[C/I]_D$ will be degraded directly by this amount. Alternatively, from Fig. 13.4,

$$[G(2°)] - [G(4°)] = 21.4 - 13.9 = \underline{\underline{7.5 \text{ dB}}}$$

It should be noted that no simple relationship can be given for calculating the effect of reduced orbital spacing on the overall $[C/I]$. The separate uplink and downlink values must be calculated and combined as described in Sec. 13.2.3. Other telecommunications authorities specify antenna characteristics that differ from the FCC specifications (see CCIR Rep. 391–3, 1978).

13.2.5 Passband interference

In the preceding section, the carrier-to-interference ratio at the receiver input is determined. However, the amount of interference reaching the detector will depend on the amount of frequency overlap between the interfering spectrum and the wanted channel passband.

Two situations can arise, as shown in Fig. 13.5. In Fig. 13.5*a*, partial overlap of the interfering signal spectra with the wanted passband is shown. The fractional interference is given as the ratio of the shaded area to the total area under the interference spectrum curve. This is denoted by Q (Sharp, 1983) or in decibels as $[Q]$. Where partial overlap occurs, Q is less than unity or $[Q] < 0$ dB. Where the interfering spectrum coincides with the wanted passband, $[Q] = 0$ dB. Evaluation of Q usually has to be carried out by computer.

The second situation, illustrated in Fig. 13.5*b*, is where multiple interfering carriers are present within the wanted passband, such as with *single carrier per channel* (SCPC) operation discussed in Sec. 14.5. Here, Q represents the sum of the interfering carrier powers within the passband, and $[Q] > 0$ dB.

In the FCC report FCC/OST R83–2 (Sharp, 1983), Q values are computed for a wide range of interfering and wanted carrier combinations.

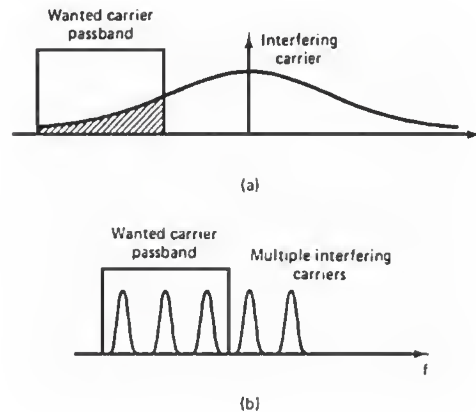


Figure 13.5 Power spectral density curves for (*a*) wideband interfering signal and (*b*) multiple interfering carriers.

Typical $[Q]$ values obtained from the FCC report are as follows: with the wanted carrier a TV/FM signal and the interfering carrier a similar TV/FM signal, $[Q] = 0$ dB; with SCPC/PSK interfering carriers, $[Q] = 27.92$ dB; and with the interfering carrier a wideband digital-type signal, $[Q] = -3.36$ dB.

The passband $[C/I]$ ratio is calculated using

$$\left[\frac{C}{I}\right]_{\mathrm{pb}} = \left[\frac{C}{I}\right]_{\mathrm{ant}} - [Q] \tag{13.7}$$

The positions of these ratios in the receiver chain are illustrated in Fig. 13.6, where it will be seen that both $[C/I]_{\mathrm{ant}}$ and $[C/I]_{\mathrm{pb}}$ are predetection ratios, measured at rf or IF. Interference also can be measured in terms of the postdetector output, shown as $[S/I]$ in Fig. 13.6, and this is discussed in the following section.

13.2.6 Receiver transfer characteristic

In some situations a measure of the interference in the postdetection baseband, rather than in the IF or rf passband, is required. Baseband interference is measured in terms of baseband signal-to-interference ratio $[S/I]$. To relate $[S/I]$ to $[C/I]_{\mathrm{ant}}$, a *receiver transfer characteristic* is introduced which takes into account the modulation characteristics of the wanted and interfering signals and the carrier frequency separation. Denoting the receiver transfer characteristic in decibels by [RTC], the relationship can be written as

$$\left[\frac{S}{I}\right] = \left[\frac{C}{I}\right]_{\mathrm{ant}} + [RTC] \tag{13.8}$$

It will be seen that [RTC] is analogous to the receiver processing gain $[G_P]$ introduced in Sec. 9.6.3. Note that it is the $[C/I]$ at the antenna which is used, not the passband value, the [RTC] taking into account any frequency offset. The [RTC] will always be a positive number of decibels so that the baseband signal-to-interference ratio will be greater than the carrier-to-interference ratio at the antenna.

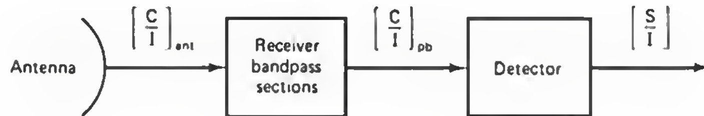


Figure 13.6 Carrier-to-interference ratios and signal-to-interference ratio.

Calculation of [RTC] for various combinations of wanted and interfering carriers is very complicated and has to be done by computer. As an example, taken from Sharp (1983), when the wanted carrier is TV/FM with a modulation index of 2.571 and the interfering carrier is TV/FM with a modulating index of 2.560, the carrier frequency separation being zero, the [RTC] is computed to be 31.70 dB. These computations are limited to low levels of interference (see Sec. 13.2.8).

13.2.7 Specified interference objectives

Although $[C/I]_{pb}$ or $[S/I]$ gives a measure of interference, ultimately, the effects of interference must be assessed in terms of what is tolerable to the end user. Such assessment usually relies on some form of subjective measurement. For TV, viewing tests are conducted, in which a mixed audience of experienced and inexperienced viewers (experienced from the point of view of assessing the effects of interference) assess the effects of interference on picture quality. By gradually increasing the interference level, a quality impairment factor can be established which ranges from 1 to 5. The five grades are defined as (Chouinard, 1984)

5. Imperceptible
4. Perceptible, but not annoying
3. Slightly annoying
2. Annoying
1. Very annoying

Acceptable picture quality requires a quality impairment factor of at least 4.2. Typical values of interference levels which result in acceptable picture quality are for broadcast TV, $[S/I] = 67$ dB; and for cable TV, $[C/I]_{pb} = 20$ dB.

For digital circuits, the $[C/I]_{pb}$ is related to the *bit error rate* (BER) (see, e.g., CCIR Rec. 523, 1978). Values of the required $[C/I]_{pb}$ used by Sharp (1983) for different types of digital circuits range from 20 to 32.2 dB.

To give some idea of the numerical values involved, a summary of the objectives stated in the FCC single-entry interference program is presented in Table 13.1. In some cases the objective used differed from the reference objective, and the values used are shown in parentheses. In some entries in the table, noise is shown measured in units of pW0p. Here, the pW stands for picowatts. The 0 means that the noise is measured at a “zero-level test point,” which is a point in the circuit where a test-tone signal produces a level of 0 dBm. The final p stands for psophometrically weighted noise, discussed in Sec. 9.6.6.

TABLE 13.1 Summary of Single-Entry Interference Objectives Used in FCC/OST R83-2, May 1983

[*S*/*I*] objectives:
FDM/FM: 600 pW0p or [*S*/*I*] = 62.2 dB (62.2 dB). Reference: CCIR Rec. 466–3.
TV/FM (broadcast quality): [*S*/*I*] = 67 dB weighted (67 dB; 65.5 dB). References: CCIR Rec. 483–1, 354–2, 567, 568.
CSSB/AM: (54.4 dB; 62.2 dB). No references quoted.
[*C*/*I*]pb objectives:
TV/FM (CATV): [*C*/*I*]pb = 20 dB (22 dB; 27 dB). Reference: ABC 62 FCC 2d 901 (1976).
Digital: [*C*/*I*]pb = [*C*/*N*] (at BER 10^{-6}) + 14 dB (20 to 32.2 dB). Reference: CCIR Rec. 523.
SCPC/PSK: (21.5 dB; 24 dB). No references quoted.
SCPC/FM: (21.2 dB; 23.2 dB). No references quoted.
SS/PSK: (11 dB; 0.6 dB). No references quoted.

13.2.8 Protection ratio

In CCIR Report 634–2 (1982), the *International Radio Consultative Committee* (CCIR) specifies the permissible interference level for TV carriers in terms of a parameter known as the *protection ratio*. The protection ratio is defined as the minimum carrier-to-interference ratio at the input to the receiver which results in "just perceptible" degradation of picture quality. The protection ratio applies only for wanted and interfering TV carriers at the same frequency, and it is equivalent to $[C/I]_{\text{pb}}$ evaluated for this situation. Denoting the quality impairment factor by Q_{IF} and the protection ratio in decibels by $[\text{PR}_0]$, the equation given in CCIR Report 634–2 (1982) is

$$[\text{PR}_0] = 12.5 - 20\log\left(\frac{\text{Dv}}{12}\right) - Q_{\text{IF}} + 1.1Q_{\text{IF}}^2 \tag{13.9}$$

Here Dv is the peak-to-peak deviation in megahertz.

Example 13.5 An FM/TV carrier is specified as having a modulation index of 2.571 and a top modulating frequency of 4.2 MHz. Calculate the protection ratio required to give a quality impairment factor of (*a*) 4.2 and (*b*) 4.5.

Solution The peak-to-peak deviation is $2 \times 2.571 \times 4.2 = 21.6$ MHz. Applying Eq. (13.9) gives the following results:

$$(a) \quad [\text{PR}_0] = 12.5 - 20\log\left(\frac{21.6}{12}\right) - 4.2 + 1.1 \times 4.2^2$$

$$= \underline{\underline{22.6 \text{ dB}}}$$

$$(b) \quad [\text{PR}_0] = 12.5 - 20\log\left(\frac{21.6}{12}\right) - 4.5 + 1.1 \times 4.5^2$$

$$= \underline{\underline{25.2 \text{ dB}}}$$

It should be noted that the receiver transfer characteristic discussed in Sec. 13.2.6 was developed from the CCIR protection ratio concept (see Jeruchim and Kane, 1970).

13.3 Energy Dispersal

The power in a frequency-modulated signal remains constant, independent of the modulation index. When unmodulated, all the power is at the carrier frequency, and when modulated, the same total power is distributed among the carrier and the sidebands. At low modulation indices the sidebands are grouped close to the carrier, and the power spectral density, or wattage per unit bandwidth, is relatively high in that spectral region. At high modulation indices, the spectrum becomes widely spread, and the power spectral density relatively low.

Use is made of this property in certain situations to keep radiation within CCIR recommended limits. For example, to limit the A_2 mode of interference in the 1- to 15-GHz range for the fixed satellite service, CCIR Radio Regulations state, in part, that the earth station EIRP should not exceed 40 dBW in any 4-kHz band for $\Theta \leq 0°$ and should not exceed $40 + 3\Theta$ dBW in any 4-kHz band for $0° < \Theta \leq 5°$. The angle Θ is the angle of elevation of the horizon viewed from the center of radiation of the earth station antenna. It is positive for angles above the horizontal plane, as illustrated in Fig. 13.7*a*, and negative for angles below the horizontal plane, as illustrated in Fig. 13.7*b*.

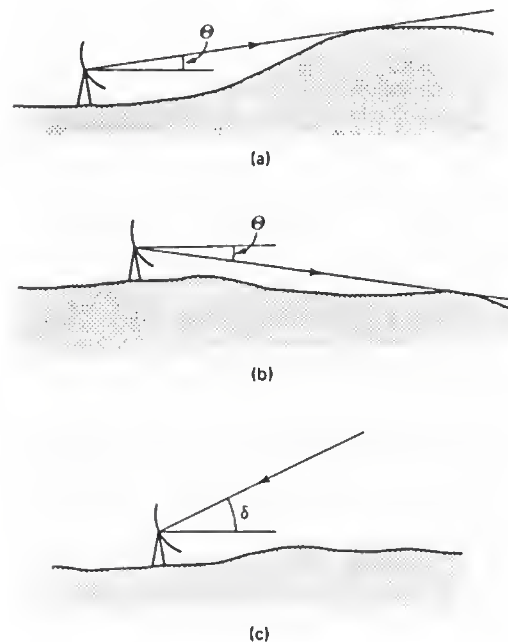


Figure 13.7 Angles θ and δ as defined in Sec. 13.3.

For space stations transmitting in the frequency range 3400 to 7750 MHz, the limits are specified in terms of power flux density for any 4- kHz bandwidth. Denoting the angle of arrival as δ degrees, measured above the horizontal plane as shown in Fig. 13.7*c*, these limits are

- -152 dB(W/m^2) in any 4-kHz band for $0° \le \delta \le 5°$
- $-152 + 0.5\,(\delta - 5)$ dB(W/m^2) in any 4-kHz band for $5° < \delta \le 25°$
- -142 dB(W/m^2) in any 4-kHz band for $25° < \delta \le 90°$

Because the specification is in terms of power or flux density in any 4-kHz band, not the total power or the total flux density, a carrier may be within the limits when heavily frequency-modulated, but the same carrier with light frequency modulation may exceed the limits. An energy-dispersal waveform is a low-frequency modulating wave which is inserted below the lowest baseband frequency for the purpose of dispersing the spectral energy when the current value of the modulating index is low. In the INTELSAT system for FDM carriers, a symmetrical triangular wave is used, a different fundamental frequency for this triangular wave in the range 20 to 150 Hz being assigned to each FDM carrier. The rms level of the baseband is monitored, and the amplitude of the dispersal waveform is automatically adjusted to keep the overall frequency deviation within defined limits. At the receive end, the dispersal waveform is removed from the demodulated signal by lowpass filtering.

With television signals the situation is more complicated. The dispersal waveform, usually a sawtooth waveform, must be synchronized with the field frequency of the video signal to prevent video interference, so for the 525/60 standard, a 30-Hz wave is used, and for the 625/50 standard, a 25-Hz wave is used. If the TV signal occupies the full bandwidth of the transponder, known as *full-transponder television*, the dispersal level is kept constant at a peak-to-peak deviation of 1 MHz irrespective of the video level. In what is termed *half-transponder television*, where the TV carrier occupies only one-half of the available transponder bandwidth, the dispersal deviation is maintained at 1-MHz peak to peak when video modulation is present and is automatically increased to 2 MHz when video modulation is absent. At the receiver, video clamping is the most commonly used method of removing the dispersal waveform.

Energy dispersal is effective in reducing all modes of interference but particularly that occurring between earth and terrestrial stations (A_2 mode) and between space and terrestrial stations (C_1). It is also effective in reducing intermodulation noise.

13.4 Coordination

When a new satellite network is in the planning stage, certain calculations have to be made to ensure that the interference levels will remain within acceptable limits. These calculations include determining the interference that will be caused by the new system and interference it will receive from other satellite networks.

In Sec. 13.2, procedures were outlined showing how interference may be calculated by taking into account modulation parameters and carrier frequencies of wanted and interfering systems. These calculations are very complex, and the CCIR uses a simplified method to determine whether *coordination* is necessary. As mentioned previously, where the potential for interference exists, the telecommunication administrations are required to coordinate the steps to be taken to reduce interference, a process referred to as *coordination*.

To determine whether or not coordination is necessary, the interference level is calculated assuming maximum spectrum density levels of the interfering signals and converted to an equivalent increase in noise temperature. The method is specified in detail in CCIR Report 454–3 (1982) for a number of possible situations. To illustrate the method, one specific situation where the existing and proposed systems operate on the same uplink and downlink frequencies, will be explained here.

Figure 13.8*a* shows the two networks, R and R'. The method will be described for network R' interfering with the operation of R. Satellite S' can interfere with the earth-station E, this being a B_1 mode of interference, and earth-station E' can interfere with the satellite S, this being a B_2 mode. Note that the networks need not be physically adjacent to one another.

13.4.1 Interference levels

Consider first the interference B_1. This is illustrated in Fig. 13.8*b*. Let U_S represent the maximum power density transmitted from satellite S'. The units for U_S are W/Hz, or joules (J), and this quantity is explained in more detail shortly. Let the transmit gain of satellite S' in the direction of earth-station E be G'_S, and let G_E be the receiving gain of earth-station E in the direction of satellite S'. The interfering spectral power density received by the earth station is therefore

$$[I_1] = [U_S] + [G'_S] + [G_E] - [L_D] \tag{13.10}$$

where L_D is the propagation loss for the downlink. The gain and loss factors are power ratios, and the brackets denote the corresponding decibel values as before. The increase in equivalent noise temperature

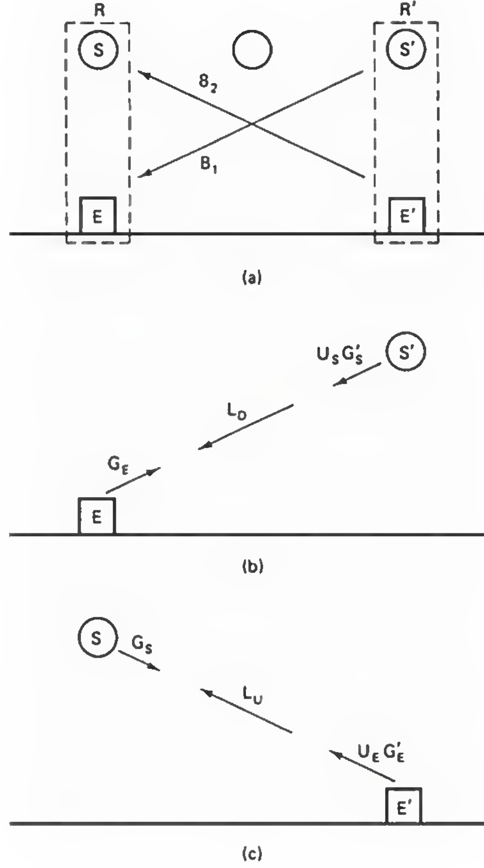


Figure 13.8 (*a*) Interference modes B_1 and B_2 from network R' into network R. (*b*) For the B_1 mode the interfering power in dBW/Hz is $[I_1] = [U_S] + [G'_S] + [G_E] - [L_D]$. (*c*) For the B_2 mode the interfering power in dBW/Hz is $[I_2] = [U_E] + [G'_E] + [G_S] - [L_U]$.

at the earth-station receiver input can then be defined using Eq. (12.15) as

$$\begin{aligned}[\Delta T_E] &= [I_1] - [k] \\ &= [I_1] + 228.6\end{aligned} \tag{13.11}$$

Here, k is Boltzmann's constant and $[k] = -228.6$ dBJ/K.

A similar argument can be applied to the uplink interference B_2, as illustrated in Fig. 13.8*c*, giving

$$[I_2] = [U_E] + [G'_E] + [G_S] - [L_U] \tag{13.12}$$

The corresponding increase in the equivalent noise temperature at the satellite receiver input is then

$$[\Delta T_S] = [I_2] + 228.6 \tag{13.13}$$

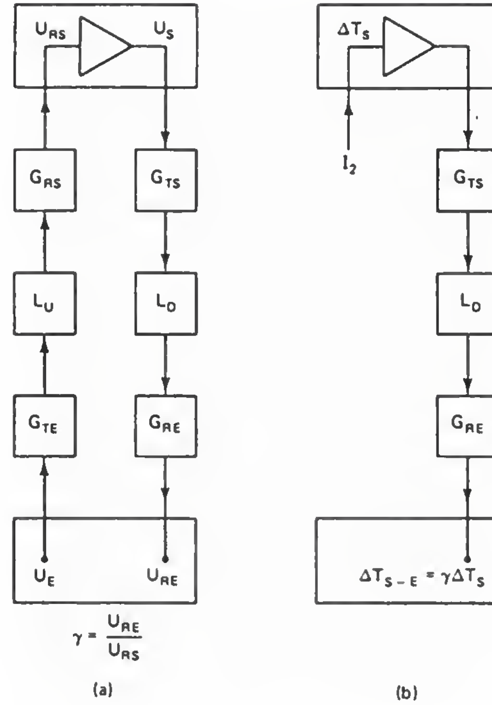


Figure 13.9 (*a*) Defining the transmission gain γ in Sec. 13.4.2. (*b*) Use of transmission gain to refer satellite noise temperature to an earth station.

Here, U_E is the maximum power spectral density transmitted by earth station E', G'_E is the transmit gain of E' in the direction of S, G_S is the receive gain of S in the direction of E', and L_U is the uplink propagation loss.

13.4.2 Transmission gain

The effect of the equivalent temperature rise ΔT_S must be transferred to the earth-station E, and this is done using the transmission gain for system R, which is calculated for the situation shown in Fig. 13.9. Figure 13.9*a* shows the satellite circuit in block schematic form. U_E represents the maximum power spectral density transmitted by earth-station E, and G_{TE} represents the transmit gain of E in direction S. G_{RS} represents the receive gain of S in direction E. The received power spectral density at satellite S is therefore

$$[U_{\mathrm{RS}}] = [U_E] + [G_{\mathrm{TE}}] + [G_{\mathrm{RS}}] - [L_U] \tag{13.14}$$

In a similar way, with satellite S transmitting and earth-station E receiving, the received power spectral density at earth-station E is

$$[U_{\mathrm{RE}}] = [U_S] + [G_{\mathrm{TS}}] + [G_{\mathrm{RE}}] - [L_D] \tag{13.15}$$

where U_S is the maximum power spectral density transmitted by S, G_{TS} is the transmit gain of S in the direction of E, and G_{RE} is the receive gain of E in the direction of S. It is assumed that the uplink and downlink propagation losses, L_U and L_D, are the same as those used for the interference signals.

The transmission gain for network R is then defined as

$$[\gamma] = [U_{RE}] - [U_{RS}] \tag{13.16}$$

Note that this is the same transmission gain shown in Fig. 12.9.

Using the transmission gain, the interference I_2 at the satellite may be referred to the earth-station receiver as γI_2, and hence the noise-temperature rise at the satellite receiver input may be referred to the earth-station receiver input as $\gamma \Delta T_S$. This is illustrated in Fig. 13.9*b*. Expressed in decibel units, the relationship is

$$[\Delta T_{S-E}] = [\gamma] + [\Delta T_S] \tag{13.17}$$

13.4.3 Resulting noise-temperature rise

The overall equivalent rise in noise temperature at earth-station E as a result of interference signals B_1 and B_2 is then

$$\Delta \mathrm{T} = \Delta T_{S-E} + \Delta T_E \tag{13.18}$$

In this final calculation the dBK values must first be converted to degrees, which are then added to give the resulting equivalent noise-temperature rise at the earth-station E receive antenna output.

Example 13.6 Given that $L_U = 200$ dB, $L_D = 196$ dB, $G_E = G'_E = 25$ dB, $G_S = G'_S = 9$ dB, $G_{TE} = G_{RE} = 48$ dB, $G_{RS} = G_{TS} = 19$ dB, $U_S = U'_S = 1\ \mu$J, and $U'_E = 10\ \mu$J; calculate the transmission gain $[\gamma]$, the interference levels $[I_1]$ and $[I_2]$, and the equivalent temperature rise overall.

Solution Using Eq. (13.14) gives

$$[U_{RS}] = -50 + 48 + 19 - 200$$
$$= -183 \text{ dBJ}$$

Using Eq. (13.15) gives

$$[U_{RE}] = -60 + 19 + 48 - 196$$
$$= -189 \text{ dBJ}$$

Therefore,

$$[\gamma] = -189 - (-183)$$
$$= -6 \text{ dB}$$

From Eq. (13.10),

$$[I_1] = -60 + 9 + 25 - 196$$
$$= -222 \text{ dBJ}$$

From Eq. (13.12),

$$[I_2] = -50 + 25 + 9 - 200$$
$$= -216 \text{ dBJ}$$

From Eq. (13.11),

$$[\Delta T_E] = -222 + 228.6$$
$$= 6.6 \text{ dBK} \quad \text{or} \quad \Delta T_E = 4.57 \text{ K}$$

From Eq. (13.13),

$$[\Delta T_S] = -216 + 228.6$$
$$= 12.6 \text{ dBK}$$

From Eq. (13.17),

$$[\Delta T_{S-E}] = -6 + 12.6$$
$$= 6.6 \text{ dBK} \quad \text{or} \quad \Delta T_{S-E} = 4.57 \text{ K}$$

The resulting equivalent noise-temperature rise at the earth-station E receive antenna output is 4.57 + 4.57 = 9.14 K.

13.4.4 Coordination criterion

CCIR Report 454–3 (1982) specifies that the equivalent noise-temperature rise should be no more than 4 percent of the equivalent thermal noise temperature of the satellite link. The equivalent thermal noise temperature is defined in the CCIR Radio Regulations, App. 29.

As an example, the CCIR Recommendations for FM Telephony allows up to 10,000 pW0p total noise in a telephone channel. The abbreviation pW0p stands for picowatts at a zero-level test point, psophometrically weighted, as already defined in connection with Table 13.1. The 10,000-pW0p total includes a 1000-pW0p allowance for terrestrial station interference and 1000 pW0p for interference from other satellite links. Thus the thermal noise allowance is 10,000 − 2000 = 8000 pW0p. Four percent of this is 320 pW0p. Assuming that this is over a 3.1-kHz bandwidth, the spectrum density is 320/3100 or approximately 0.1 pJ0p (pW0p/Hz). In decibels, this is −130 dBJ. This is output noise, and to

relate it back to the noise temperature at the antenna, the overall gain of the receiver from antenna to output, including the processing gain, discussed in Sec. 9.6.3, must be known. For illustration purposes, assume that the gain is 90 dB, so the antenna noise is $-130 - 90 = -220$ dBJ. The noise-temperature rise corresponding to this is $-220 + 228.6 = 8.6$ dBK. Converting this to kelvins gives 7.25 K.

13.4.5 Noise power spectral density

The concept of noise power spectral density was introduced in Sec. 12.5 for a flat frequency spectrum. Where the spectrum is not flat, an average value for the spectral density can be calculated. To illustrate this, the very much simplified spectrum curve of Fig. 13.10 will be used. The maximum spectrum density is flat at 3 W/Hz from 0 to 2 kHz and then slopes linearly down to zero over the range from 2 to 8 kHz.

The noise power in any given bandwidth is calculated as the area under the curve, the width of which is the value of the bandwidth. Thus, for the first 2 kHz, the noise power is 3 W/Hz $\times$ 2000 = 6000 W. From 2 to 8 kHz, the noise power is $3 \times (8 - 2) \times 1000/2 = 9000$ W. The total power is therefore 15,000 W, and the average spectral density is 15,000/8000 = 1.875 W/Hz.

The noise power spectral density over the worst 4-kHz bandwidth must include the highest part of the curve and is therefore calculated for the 0- to 4-kHz band. The power over this band is seen to be the area of the rectangle 3 W/Hz $\times$ 4 kHz minus the area of the triangle shown dashed in Fig. 13.10. The power over the 0- to 4-kHz band is therefore $(3 \times 4000) - (3 - 2) \times (4 - 2) \times 1000/2 = 11{,}000$ W, and the spectral density is 11,000/4000 = 2.75 W/Hz.

The units for spectral power density are often stated as watts per hertz (W/Hz). Expressed in this manner the units are descriptive of the way in which the power spectral density is arrived at. In terms of fundamental units, watts are equivalent to joules per second and hertz to cycles per second or simply seconds^{-1}, since cycles are a dimensionless

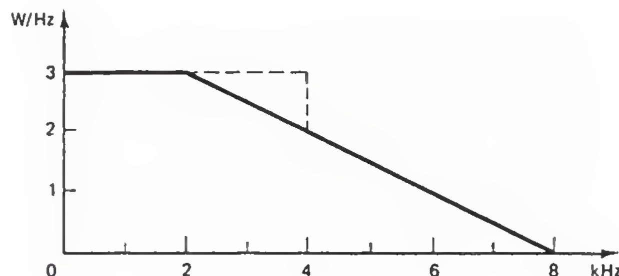


Figure 13.10 Power spectrum density curve (see Sec. 13.4.5).

quantity. Thus, 1 W/Hz is equivalent to 1 J/s $\div$ s^{-1}, which is simply J. The units for power spectral density therefore can be stated as joules. This is verified by Eq. (12.15) for noise power spectral density.

13.5 Problems and Exercises

13.1. Describe briefly the modes of interference that can occur in a satellite communications system. Distinguish carefully between satellite and terrestrial modes of interference.

13.2. Define and explain the difference between topocentric angles and geocentric angles as applied to satellite communications. Two geostationary satellites have an orbital spacing of 4°. Calculate the topocentric angle subtended by the satellites, measured (*a*) from the midpoint between the subsatellite points and (*b*) from either of the subsatellite points.

13.3. Westar IV is located at 98.5°W and Telstar at 96°W. The coordinates for two earth stations are 104°W, 36°N and 90°W, 32°N. By using the look angle and range formulas given in Sec. 3.2, calculate the topocentric angle subtended at each earth station by these two satellites.

13.4. Explain what is meant by *single-entry interference.* Explain why it is the radiation pattern of the earth-station antennas, not the satellite antennas, which governs the level of interference.

13.5. A geostationary satellite employs a 3.5-m parabolic antenna at a frequency of 12 GHz. Calculate the −3-dB beamwidth and the spot diameter on the equator.

13.6. Calculate the −3-dB beamwidth for an earth-station antenna operating at 14 GHz. The antenna utilizes a parabolic reflector of 3.5-m diameter. Compare the distance separation of satellites at 2° spacing with the diameter of the beam at the −3-dB points on the geostationary orbit.

13.7. Compare the increase in interference levels expected when satellite orbital spacing is reduced from 4° to 2° for earth-station antenna sidelobe patterns of (*a*) $32 - 25 \log \theta$ dB and (*b*) $29 - 25 \log \theta$ dB.

13.8. A satellite circuit operates with an uplink transmit power of 28.3 dBW and an antenna gain of 62.5 dB. A potential interfering circuit operates with an uplink power of 26.3 dBW. Assuming a 4-dB polarization discrimination figure and earth-station sidelobe gain function of $32 - 25 \log \theta$ dB, calculate the [C/I] ratio at the satellite for 2° satellite spacing.

13.9. The downlink of a satellite circuit operates at a satellite [EIRP] of 35 dBW and a receiving earth-station antenna gain of 59.5 dB. Interference is produced by a satellite spaced 3°, its [EIRP] also being 35 dBW. Calculate the [C/I] ratio

at the receiving antenna, assuming 6-dB polarization discrimination. The sidelobe gain function for the earth-station antenna is $32 - 25 \log \theta$.

13.10. A satellite circuit operates with an earth-station transmit power of 30 dBW and a satellite [EIRP] of 34 dBW. This circuit causes interference with a neighboring circuit for which the earth-station transmit power is 24 dBW, the transmit antenna gain is 54 dB, the satellite [EIRP] is 34 dBW, and the receive earth-station antenna gain is 44 dB. Calculate the carrier-to-interference ratio at the receive station antenna. Antenna sidelobe characteristics of $32 - 25 \log \theta$ dB may be assumed, with the satellites spaced at 2° and polarization isolation of 4 dB on uplink and downlink.

13.11. Repeat Prob. 13.10 for a sidelobe pattern of $29 - 25 \log \theta$.

13.12. A satellite TV/FM circuit operates with an uplink power of 30 dBW, an antenna gain of 53.5 dB, and a satellite [EIRP] of 34 dBW. The destination earth station has a receiver antenna gain of 44 dB. An interfering circuit has an uplink power of 11.5 dBW and a satellite [EIRP] of 15.7 dBW. Given that the spectral overlap is $[Q] = 3.8$ dB, calculate the passband $[C/I]$ ratio. Assume polarization discrimination figures of 4 dB on the uplink and 0 dB on the downlink and an antenna sidelobe pattern of $32 - 25 \log \theta$. The satellite spacing is 2°.

13.13. Repeat Prob. 13.12 for an interfering carrier for which the uplink transmit power is 26 dBW, the satellite [EIRP] is 35.7 dBW, and the spectral overlap figure is −3.36 dB.

13.14. An FDM/FM satellite circuit operates with an uplink transmit power of 11.9 dBW, an antenna gain of 53.5 dB, and a satellite [EIRP] of 19.1 dBW. The destination earth station has an antenna gain of 50.5 dB. A TV/FM interfering circuit operates with an uplink transmit power of 28.3 dBW and a satellite [EIRP] of 35 dBW. Polarization discrimination figures are 6 dB on the uplink and 0 dB on the downlink. Given that the receiver transfer characteristic for wanted and interfering signals is [RTC] = 60.83 dB, calculate the baseband $[S/I]$ ratio for an antenna sidelobe pattern of $29 - 25 \log \theta$. The satellite spacing is 2°.

13.15. Repeat Prob. 13.14 for an interfering circuit operating with an uplink transmit power of 27 dBW, a satellite [EIRP] of 34.2 dBW, and [RTC] = 37.94 dB.

13.16. Explain what is meant by *single-entry interference objectives.* Show that an interference level of 600 pW0p is equivalent to an $[S/I]$ ratio of 62.2 dB.

13.17. For the wanted circuit in Probs. 13.12 and 13.13, the specified interference objective is $[C/I]_{pb} = 22$ dB. Is this objective met?

13.18. For broadcast TV/FM, the permissible video-to-noise objective is specified as $[S/N] = 53$ dB. The CCIR recommendation for interference is that the total interference from all other satellite networks should not exceed 10 percent of the video noise and that the single-entry interference should not exceed 40 percent of this total. Show that this results in a single-entry objective of $[S/I] = 67$ dB.

13.19. Explain what is meant by *protection ratio*. A TV/FM carrier operates at a modulation index of 2.619, the top modulating frequency being 4.2 MHz. Calculate the protection ratio required for quality factors of (*a*) 4.2 and (*b*) 4.5. How do these values compare with the specified interference objective of $[C/I]_{pb} = 22$ dB?

13.20. Explain what is meant by *energy dispersal* and how this may be achieved.

13.21. Explain what is meant by *coordination* in connection with interference assessment in satellite circuits.

References

CCIR Recommendation 523. 1978. "Maximum Permissible Levels of Interference in a Geostationary Satellite Network in the Fixed Satellite Service Using 8-bit PCM Encoded Telephony Caused by Other Networks of This Service." *14th Plenary Assembly*, Vol. IV, Kyoto.

CCIR Report 391-3. 1978. "Radiation Diagrams of Antennae for Earth Stations in the Fixed Satellite Service for Use in Interference Studies and for the Determination of a Design Objective." *14th Plenary Assembly*, Vol. IV, Kyoto.

CCIR Report 454-3. 1982. "Method of Calculation to Determine Whether Two Geostationary-Satellite Systems Require Coordination." *15th Plenary Assembly*, Vol. IX, Part 1, Geneva.

CCIR Report 634-2. 1982. "Maximum Interference Protection Ratio for Planning Television Broadcast Systems." *Broadcast Satellite Service (Sound and Television)*, Vols. X and IX, Part 2, Geneva.

Chouinard, G. 1984. "The Implications of Satellite Spacing on TVRO Antennas and DBS Systems." *Canadian Satellite User Conference*, Ottawa.

ITU. 1985. *Handbook on Satellite Communications* (FSS). Geneva.

ITU. 1986. *Radio Regulations*. Geneva.

Jeruchim, M. C., and D. A. Kane. 1970. *Orbit/Spectrum Utilization Study*, Vol. IV. General Electric Doc. No. 70SD4293, December 31.

Microwave Filter Company, Inc. 1984. "TI and TVROs: A Brief Troubleshooting Guide to Suppressing Terrestrial Interference at 3.7–4.2 GHz TVRO Earth Stations."

Sharp, G. L. 1983. Reduced Domestic Satellite Orbital Spacings at 4/6 GHz. FCC/OST R83-2, May.

Sharp, G. L. 1984*a*. "Reduced Domestic Satellite Orbit Spacing." *AIAA Communications Satellite Systems Conference*, Orlando, FL, March 18–22.

Chapter

14

Satellite Access

14.1 Introduction

A transponder channel aboard a satellite may be fully loaded by a single transmission from an earth station. This is referred to as a *single access* mode of operation. It is also possible, and more common, for a transponder to be loaded by a number of carriers. These may originate from a number of earth stations geographically separate, and each earth station may transmit one or more of the carriers. This mode of operation is termed *multiple access*. The need for multiple access arises because more than two earth stations, in general, will be within the service area of a satellite. Even so-called spot beams from satellite antennas cover areas several hundred miles across.

The two most commonly used methods of multiple access are *frequency-division multiple access* (FDMA) and *time-division multiple access* (TDMA). These are analogous to frequency-division multiplexing (FDM) and time-division multiplexing (TDM) described in Chaps. 9 and 10. However, multiple access and multiplexing are different concepts, and as pointed out in CCIR Report 708 (1982), modulation (and hence multiplexing) is essentially a transmission feature, whereas multiple access is essentially a traffic feature.

A third category of multiple access is *code-division multiple access* (CDMA). In this method each signal is associated with a particular code that is used to spread the signal in frequency and/or time. All such signals will be received simultaneously at an earth station, but by using the key to the code, the station can recover the desired signal by means of correlation. The other signals occupying the transponder channel appear very much like random noise to the correlation decoder.

Multiple access also may be classified by the way in which circuits are assigned to users (*circuits* in this context implies one communication

channel through the multiple-access transponder). Circuits may be *preassigned*, which means they are allocated on a fixed or partially fixed basis to certain users. These circuits are therefore not available for general use. Preassignment is simple to implement but is efficient only for circuits with *continuous heavy* traffic.

An alternative to preassignment is *demand-assigned multiple access* (DAMA). In this method, all circuits are available to all users and are assigned according to the demand. DAMA results in more efficient overall use of the circuits but is more costly and complicated to implement.

Both FDMA and TDMA can be operated as preassigned or demand assigned systems. CDMA is a random-access system, there being no control over the timing of the access or of the frequency slots accessed.

These multiple-access methods refer to the way in which a single *transponder* channel is utilized. A satellite carries a number of transponders, and normally each covers a different frequency channel, as shown in Fig. 7.13. This provides a form of FDMA to the whole satellite. It is also possible for transponders to operate at the same frequency but to be connected to different spot-beam antennas. These allow the satellite as a whole to be accessed by earth stations widely separated geographically but transmitting on the same frequency. This is termed *frequency reuse*. This method of access is referred to as *space-division multiple access* (SDMA). It should be kept in mind that each spot beam may itself be carrying signals in one of the other multiple-access formats.

14.2 Single Access

With single access, a single modulated carrier occupies the whole of the available bandwidth of a transponder. Single-access operation is used on heavy-traffic routes and requires large earth station antennas such as the class A antenna shown in Fig. 8.7. As an example, Telesat Canada provides heavy route message facilities, with each transponder channel being capable of carrying 960 one-way voice circuits on an FDM/FM carrier, as illustrated in Fig. 14.1. The earth station employs a 30-m-diameter

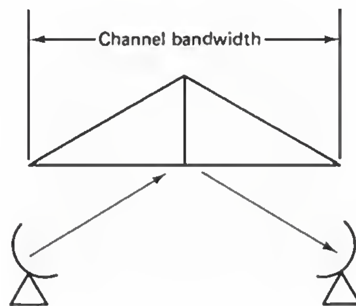


Figure 14.1 Heavy route message (frequency modulation—single access). (*Courtesy of Telesat Canada, 1983.*)

antenna and a parametric amplifier, which together provide a minimum $[G/T]$ of 37.5 dB/K.

14.3 Preassigned FDMA

Frequency slots may be preassigned to analog and digital signals, and to illustrate the method, analog signals in the FDM/FM/FDMA format will be considered first. As the acronyms indicate, the signals are frequency-division multiplexed, frequency modulated (FM), with FDMA to the satellite. In Chap. 9, FDM/FM signals are discussed. It will be recalled that the voice-frequency (telephone) signals are first SSBSC amplitude modulated onto voice carriers in order to generate the single sidebands needed for the FDM. For the purpose of illustration, each earth station will be assumed to transmit a 60-channel supergroup. Each 60-channel supergroup is then frequency modulated onto a carrier which is then upconverted to a frequency in the satellite uplink band.

Figure 14.2 shows the situation for three earth stations: one in Ottawa, one in New York, and one in London. All three earth stations

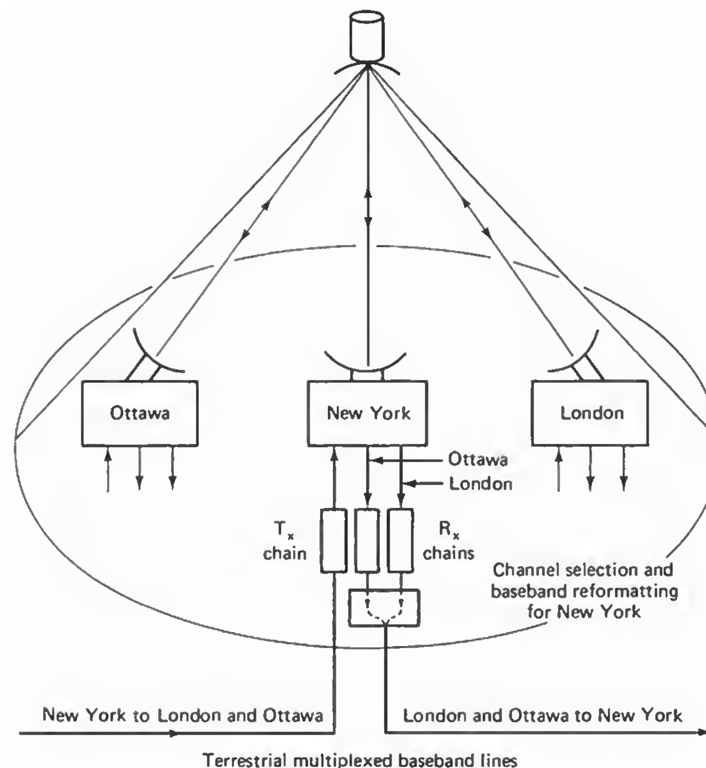


Figure 14.2 Three earth stations transmitting and receiving simultaneously through the same satellite transponder, using fixed-assignment FDMA.

access a single satellite transponder channel simultaneously, and each communicates with both of the others. Thus it is assumed that the satellite receive and transmit antenna beams are *global*, encompassing all three earth stations. Each earth station transmits one uplink carrier modulated with a 60-channel supergroup and receives two similar downlink carriers.

The earth station at New York is shown in more detail. One transmit chain is used, and this carries telephone traffic for both Ottawa and London. On the receive side, two receive chains must be provided, one for the Ottawa-originated carrier and one for the London-originated carrier. Each of these carriers will have a mixture of traffic, and in the demultiplexing unit, only those telephone channels intended for New York are passed through. These are remultiplexed into an FDM/FM format which is transmitted out along the terrestrial line to the New York switching office. This earth-station arrangement should be compared with that shown in Fig. 8.6.

Figure 14.3 shows a hypothetical frequency assignment scheme for the hypothetical network of Fig. 14.2. Uplink carrier frequencies of 6253, 6273, and 6278 MHz are shown for illustration purposes. For the satellite transponder arrangement of Fig. 7.13, these carriers would be translated down to frequencies of 4028, 4048, and 4053 MHz (i.e., the corresponding 4-GHz-band downlink frequencies) and sent to transponder 9 of the satellite. Typically, a 60-channel FDM/FM carrier occupies 5 MHz of transponder bandwidth, including guardbands. A total frequency

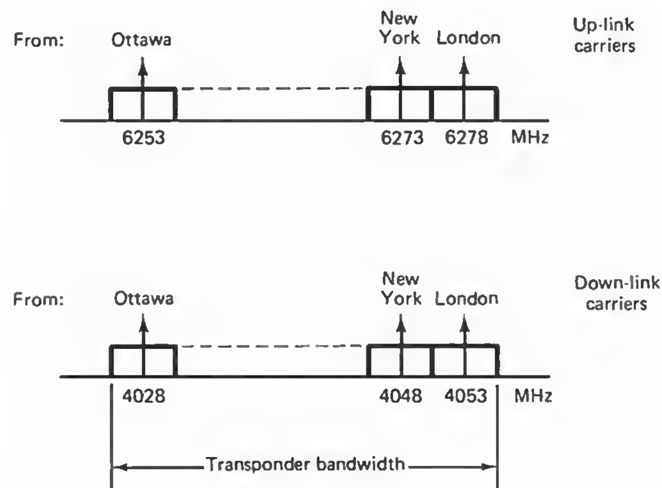


Figure 14.3 Transponder channel assignments for the earth stations shown in Fig. 14.2.

allowance of 15 MHz is therefore required for the three stations, and each station receives all the traffic. The remainder of the transponder bandwidth may be unused, or it may be occupied by other carriers, which are not shown.

As an example of preassignment, suppose that each station can transmit up to 60 voice circuits and that 40 of these are preassigned to the New York–London route. If these 40 circuits are fully loaded, additional calls on the New York–London route will be blocked even though there may be idle circuits on the other preassigned routes.

Telesat Canada operates medium-route message facilities utilizing FDM/FM/FDMA. Figure 14.4 shows how five carriers may be used to support 168 voice channels. The earth station that carries the full load has a [*G*/*T*] of 37.5 dB/K, and the other four have [*G*/*T*]'s of 28 dB/K.

Preassignment also may be made on the basis of a *single channel per carrier* (SCPC). This refers to a single voice (or data) channel per carrier, not a transponder channel, which may in fact carry some hundreds of voice channels by this method. The carriers may be frequency modulated or phase-shift modulated, and an earth station may be capable of transmitting one or more SCPC signals simultaneously.

Figure 14.5 shows the INTELSAT SCPC channeling scheme for a 36-MHz transponder. The transponder bandwidth is subdivided into 800 channels each 45-kHz wide. The 45 kHz, which includes a guardband, is required for each digitized voice channel, which utilizes *quadrature phase-shift keying* (QPSK) modulation. The channel information

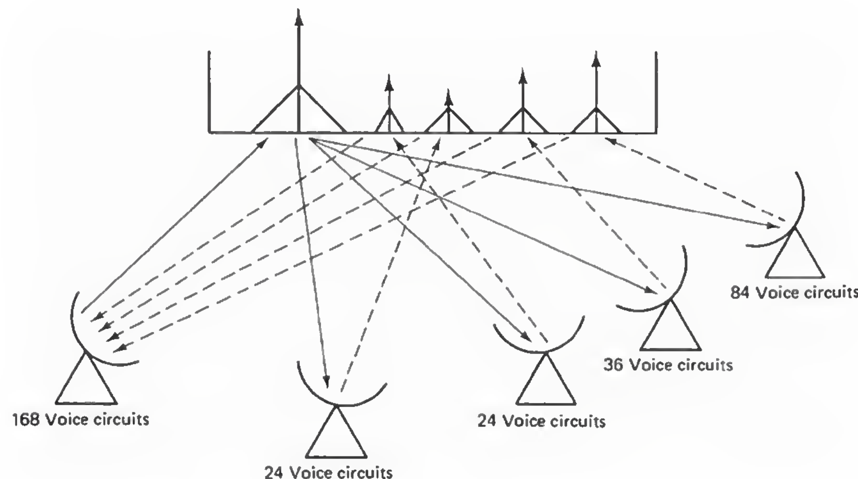


Figure 14.4 Medium route message traffic (frequency-division multiple access, FM/FDMA). (*Courtesy of Telesat Canada, 1983.*)

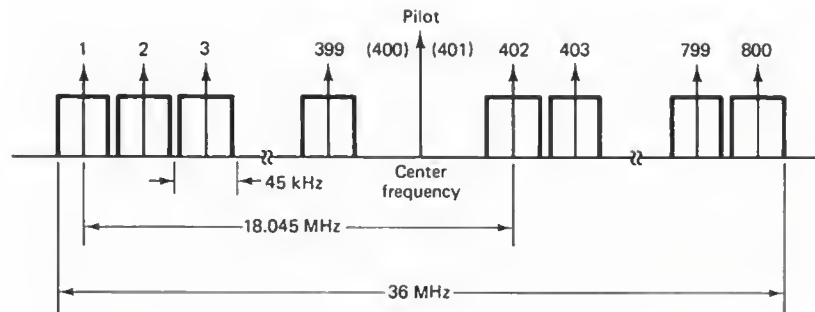


Figure 14.5 Channeling arrangement for Intelsat SCPC system.

signal may be digital data or PCM voice signals (see Chap. 10). A pilot frequency is transmitted for the purpose of frequency control, and the adjacent channel slots on either side of the pilot are left vacant to avoid interference. The scheme therefore provides a total of 798 one-way channels or up to 399 full-duplex voice circuits. In duplex operation, the frequency pairs are separated by 18.045 MHz, as shown in Fig. 14.5.

The frequency tolerance relative to the assigned values is within ±1 kHz for the received SCPC carrier and must be within ±250 Hz for the transmitted SCPC carrier (Miya, 1981). The pilot frequency is transmitted by one of the earth stations designated as a primary station. This provides a reference for *automatic frequency control* (AFC) (usually through the use of phase-locked loops) of the transmitter frequency synthesizers and receiver local oscillators. In the event of failure of the primary station, the pilot frequency is transmitted from a designated backup station.

An important feature of the INTELSAT SCPC system is that each channel is voice-activated. This means that on a two-way telephone conversation, only one carrier is operative at any one time. Also, in long pauses between speech, the carriers are switched off. It has been estimated that for telephone calls, the one-way utilization time is 40 percent of the call duration. Using voice activation, the average number of carriers being amplified at any one time by the transponder traveling-wave tube (TWT) is reduced. For a given level of intermodulation distortion (see Secs. 7.7.3 and 12.10), the TWT power output per FDMA carrier therefore can be increased.

SCPC systems are widely used on lightly loaded routes, this type of service being referred to as a *thin route service*. It enables remote earth stations in sparsely populated areas to connect into the national telephone network in a reasonably economical way. A main earth station is used to make the connection to the telephone network, as illustrated

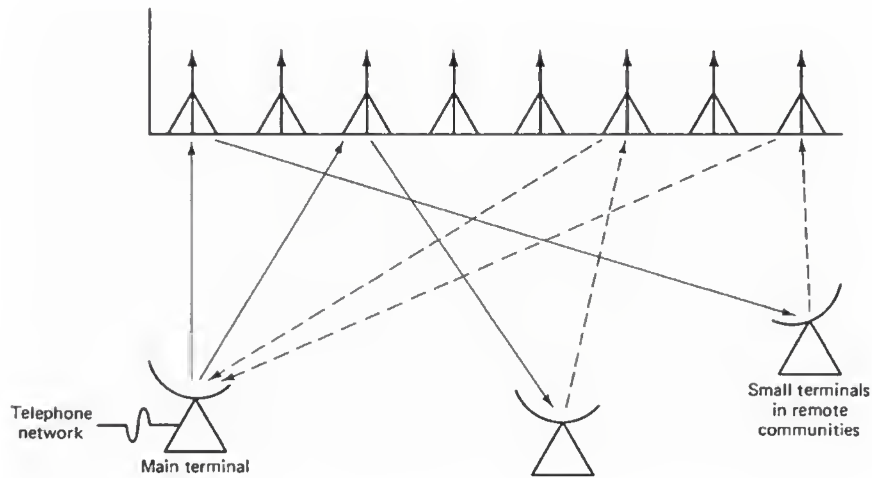


Figure 14.6 Thin route message traffic (single channel per carrier, SCPC/FDMA). (*Courtesy of Telesat Canada, 1983.*)

in Fig. 14.6. The Telesat Canada Thin Route Message Facilities provide up to 360 two-way circuits using PSK/SCPC (PSK = *phase-shift keying*). The remote terminals operate with 4.6-m-diameter antennas with [*G*/*T*] values of 19.5 or 21 dB/K. Transportable terminals are also available, one of these being shown in Fig. 14.7. This is a single-channel station that uses a 3.6-m antenna and comes complete with a desktop electronics package which can be installed on the customers' premises.



Figure 14.7 Transportable message station. (*Courtesy of Telesat Canada, 1983.*)

14.4 Demand-Assigned FDMA

In the demand-assigned mode of operation, the transponder frequency bandwidth is subdivided into a number of channels. A channel is assigned to each carrier in use, giving rise to the single-channel-per-carrier mode of operation discussed in the preceding section. As in the preassigned access mode, carriers may be frequency modulated with analog information signals, these being designated FM/SCPC, or they may be phase modulated with digital information signals, these being designated as PSK/SCPC.

Demand assignment may be carried out in a number of ways. In the polling method, a master earth station continuously polls all the earth stations in sequence, and if a *call request* is encountered, frequency slots are assigned from the pool of available frequencies. The polling delay with such a system tends to become excessive as the number of participating earth stations increases.

Instead of using a polling sequence, earth stations may request calls through the master earth station as the need arises. This is referred to as *centrally controlled random access*. The requests go over a digital orderwire, which is a narrowband digital radio link or a circuit through a satellite transponder reserved for this purpose. Frequencies are assigned, if available, by the master station, and when the call is completed, the frequencies are returned to the pool. If no frequencies are available, the blocked call requests may be placed in a queue, or a second call attempt may be initiated by the requesting station.

As an alternative to centrally controlled random access, control may be exercised at each earth station, this being known as *distributed control random access*. A good illustration of such a system is provided by the Spade system operated by INTELSAT on some of its satellites.

This is described in the following section.

14.5 Spade System

The word *Spade* is a loose acronym for SCPC pulse-code-modulated multiple-access demand-assignment equipment. Spade was developed by Comsat for use on the INTELSAT satellites (see, e.g., Martin, 1978) and is compatible with the INTELSAT SCPC preassigned system described in Sec. 14.3. However, the distributed-demand assignment facility requires a *common signaling channel* (CSC). This is shown in Fig. 14.8. The CSC bandwidth is 160 kHz, and its center frequency is 18.045 MHz below the pilot frequency, as shown in Fig. 14.8. To avoid interference with the CSC, voice channels 1 and 2 are left vacant, and to maintain duplex matching, the corresponding channels 1′ and 2′ are also left vacant. Recalling from Fig. 14.5 that channel 400 also must be left vacant, this requires that channel 800 be left vacant for duplex

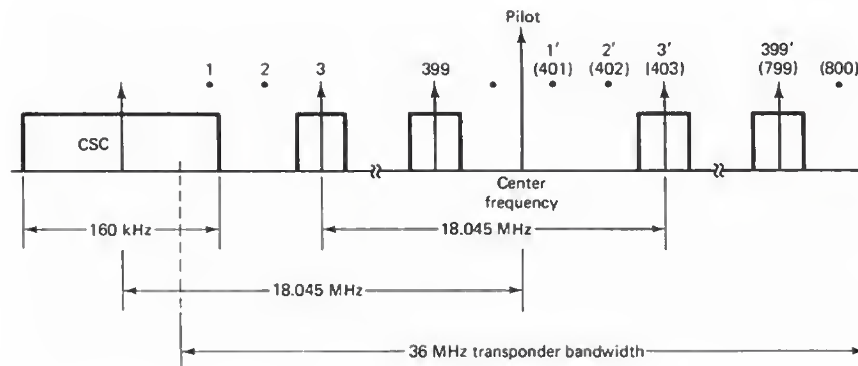


Figure 14.8 Channeling scheme for the Spade system.

matching. Thus six channels are removed from the total of 800, leaving a total of 794 one-way or 397 full-duplex voice circuits, the frequencies in any pair being separated by 18.045 MHz, as shown in Fig. 14.8. (An alternative arrangement is shown in Freeman, 1981.)

All the earth stations are permanently connected through the CSC. This is shown diagrammatically in Fig. 14.9 for six earth stations *A, B,*

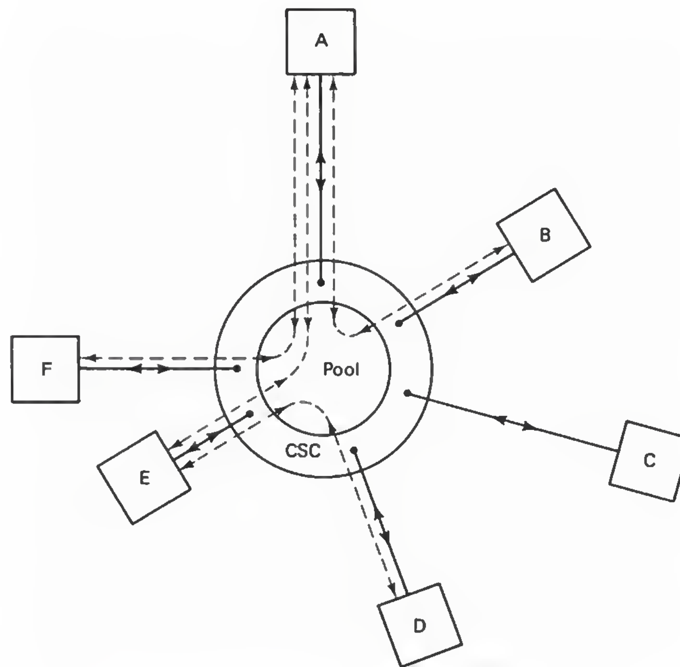


Figure 14.9 Diagrammatic representation of a Spade communications system.

C, D, E, and *F*. Each earth station has the facility for generating any one of the 794 carrier frequencies using frequency synthesizers. Furthermore, each earth station has a memory containing a list of the frequencies currently available, and this list is continuously updated through the CSC. To illustrate the procedure, suppose that a call to station *F* is initiated from station *C* in Fig. 14.9. Station *C* will first select a frequency pair at random from those currently available on the list and signal this information to station *F* through the CSC. Station *F* must acknowledge, through the CSC, that it can complete the circuit. Once the circuit is established, the other earth stations are instructed, through the CSC, to remove this frequency pair from the list.

The round-trip time between station *C* initiating the call and station *F* acknowledging it is about 600 ms. During this time, the two frequencies chosen at station *C* may be assigned to another circuit. In this event, station *C* will receive the information on the CSC update and will immediately choose another pair at random, even before hearing from station *F*.

Once a call has been completed and the circuit disconnected, the two frequencies are returned to the pool, the information again being transmitted through the CSC to all the earth stations.

As well as establishing the connection through the satellite, the CSC passes signaling information from the calling station to the destination station, in the example above from station *C* to station *F*. Signaling information in the Spade system is routed through the CSC rather than being sent over a voice channel. Each earth station has an equipment called the *demand assignment signaling and switching* (DASS) unit which performs the functions required by the CSC.

Some type of multiple access to the CSC must be provided for all the earth stations using the Spade system. This is quite separate from the SCPC multiple access of the network's voice circuits. TDMA, described in Sec. 14.7.8, is used for this purpose, allowing up to 49 earth stations to access the common signaling channel.

14.6 Bandwidth-Limited and Power-Limited TWT Amplifier Operation

A transponder will have a total bandwidth B_{TR}, and it is apparent that this can impose a limitation on the number of carriers that can access the transponder in an FDMA mode. For example, if there are K carriers each of bandwidth B, then the best that can be achieved is $K = B_{TR}/B$. Any increase in the transponder EIRP will not improve on this, and the system is said to be *bandwidth-limited*. Likewise, for digital systems, the bit rate is determined by the bandwidth, which again will be limited to some maximum value by B_{TR}.

Power limitation occurs where the EIRP is insufficient to meet the [C/N] requirements, as shown by Eq. (12.34). The signal bandwidth will be approximately equal to the noise bandwidth, and if the EIRP is below a certain level, the bandwidth will have to be correspondingly reduced to maintain the [C/N] at the required value. These limitations are discussed in more detail in the following two sections.

14.6.1 FDMA downlink analysis

To see the effects of output backoff which results with FDMA operation, consider the overall carrier-to-noise ratio as given by Eq. (12.62). In terms of noise power rather than noise power density, Eq. (12.62) states

$$\left(\frac{N}{C}\right) = \left(\frac{N}{C}\right)_U + \left(\frac{N}{C}\right)_D + \left(\frac{N}{C}\right)_{\mathrm{IM}} \tag{14.1}$$

A certain value of carrier-to-noise ratio will be needed, as specified in the system design, and this will be denoted by the subscript REQ. The overall C/N must be at least as great as the required value, a condition which can therefore be stated as

$$\left(\frac{N}{C}\right)_{\mathrm{REQ}} \geq \left(\frac{N}{C}\right) \tag{14.2}$$

Note that because the noise-to-carrier ratio rather than the carrier-to-noise ratio is involved, the actual value is equal to or less than the required value. Using Eq. (14.1), the condition can be rewritten as

$$\left(\frac{N}{C}\right)_{\mathrm{REQ}} \geq \left(\frac{N}{C}\right)_U + \left(\frac{N}{C}\right)_D + \left(\frac{N}{C}\right)_{\mathrm{IM}} \qquad 14.3)$$

The right-hand side of Eq. (14.3) is usually dominated by the downlink ratio. With FDMA, backoff is utilized to reduce the intermodulation noise to an acceptable level, and as shown in Sec. 12.10, the uplink noise contribution is usually negligible. Thus the expression can be approximated by

$$\left(\frac{N}{C}\right)_{\mathrm{REQ}} \geq \left(\frac{N}{C}\right)_D$$

or

$$\left(\frac{C}{N}\right)_{\mathrm{REQ}} \leq \left(\frac{C}{N}\right)_D \tag{14.4}$$

Consider the situation where each carrier of the FDMA system occupies a bandwidth B and has a downlink power denoted by $[\text{EIRP}]_D$. Equation (12.54) gives

$$\left[\frac{C}{N}\right]_D = [\text{EIRP}]_D + \left[\frac{G}{T}\right]_D - [\text{LOSSES}]_D - [k] - [B] \tag{14.5}$$

where it is assumed that $B_N \approx B$. This can be written in terms of the required carrier-to-noise ratio as

$$\left[\frac{C}{N}\right]_{\text{REQ}} \leq [\text{EIRP}]_D + \left[\frac{G}{T}\right]_D - [\text{LOSSES}]_D - [k] - [B] \tag{14.6}$$

To set up a reference level, consider first single-carrier operation. The satellite will have a saturation value of EIRP and a transponder bandwidth B_{TR}, both of which are assumed fixed. With single-carrier access, no backoff is needed, and Eq. (14.6) becomes

$$\left[\frac{C}{N}\right]_{\text{REQ}} \leq [\text{EIRP}_S] + \left[\frac{G}{T}\right]_D - [\text{LOSSES}]_D - [k] - [B_{\text{TR}}] \tag{14.7}$$

or

$$\left[\frac{C}{N}\right]_{\text{REQ}} - [\text{EIRP}_S] - \left[\frac{G}{T}\right]_D + [\text{LOSSES}]_D + [k] + [B_{\text{TR}}] \leq 0 \tag{14.8}$$

If the system is designed for single-carrier operation, then the equality sign applies and the reference condition is

$$\left[\frac{C}{N}\right]_{\text{REQ}} - [\text{EIRP}_S] - \left[\frac{G}{T}\right]_D + [\text{LOSSES}]_D + [k] + [B_{\text{TR}}] = 0 \tag{14.9}$$

Consider now the effect of power limitation imposed by the need for backoff. Suppose the FDMA access provides for K carriers which share the output power equally, and each requires a bandwidth B. The output power for each of the FDMA carriers is

$$[\text{EIRP}]_D = [\text{EIRP}_S] - [\text{BO}]_0 - [K] \tag{14.10}$$

The transponder bandwidth B_{TR} will be shared between the carriers, but not all of B_{TR} can be utilized because of the power limitation. Let α represent the fraction of the total bandwidth actually occupied, such that $KB = \alpha B_{\text{TR}}$, or in terms of decilogs

$$[B] = [\alpha] + [B_{\text{TR}}] - [K] \tag{14.11}$$

Substituting these relationships in Eq. (14.6) gives

$$\left[\frac{C}{N}\right]_{\mathrm{REQ}} \leq [\mathrm{EIRP}_S] - [\mathrm{BO}]_0 + \left[\frac{G}{T}\right]_D - [\mathrm{LOSSES}]_D - [k] - [B_{\mathrm{TR}}] - [\alpha] \tag{14.12}$$

It will be noted that the $[K]$ term cancels out. The expression can be rearranged as

$$\left[\frac{C}{N}\right]_{\mathrm{REQ}} - [\mathrm{EIRP}_S] - \left[\frac{G}{T}\right]_D + [\mathrm{LOSSES}]_D + [k] + [B_{\mathrm{TR}}] \leq -[\mathrm{BO}]_0 - [\alpha] \tag{14.13}$$

But as shown by Eq. (14.9), the left-hand side is equal to zero if the single carrier access is used as reference, and hence

$$0 \leq -[\mathrm{BO}]_0 - [\alpha] \qquad \text{or} \qquad [\alpha] \leq -[\mathrm{BO}]_0 \tag{14.14}$$

The best that can be achieved is to make $[\alpha] = -[\mathrm{BO}]_0$, and since the backoff is a positive number of decibels, $[\alpha]$ must be negative, or equivalently, α is fractional. The following example illustrates the limitation imposed by backoff.

Example 14.1 A satellite transponder has a bandwidth of 36 MHz and a saturation EIRP of 27 dBW. The earth-station receiver has a $[G/T]$ of 30 dB/K, and the total link losses are 196 dB. The transponder is accessed by FDMA carriers each of 3-MHz bandwidth, and 6-dB output backoff is employed. Calculate the downlink carrier-to-noise ratio for single-carrier operation and the number of carriers which can be accommodated in the FDMA system. Compare this with the number which could be accommodated if no backoff were needed. The carrier-to-noise ratio determined for single-carrier operation may be taken as the reference value, and it may be assumed that the uplink noise and intermodulation noise are negligible.

Solution Transponder bandwidth: $B_{\mathrm{TR}} = 36$ MHz, therefore $[B_{\mathrm{TR}}] = 75.56$ dBHz
Carrier bandwidth: $B = 3$ MHz, therefore $[B] = 64.77$ dBHz

For single carrier operation B_{TR} can replace B in Eq. (14.5) and $[\mathrm{EIRP}_S]$ can replace $[\mathrm{EIRP}]_D$ to give

$$\begin{aligned}\left[\frac{C}{N}\right]_D &= [\mathrm{EIRP}_S] + \left[\frac{G}{T}\right]_D - [\mathrm{LOSSES}]_D - [k] - [B_{\mathrm{TR}}] \\ &= 27 + 30 - 196 + 228.6 - 75.56 \\ &\cong \underline{\underline{14 \text{ dB}}}\end{aligned}$$

Setting $[\alpha] = -[\mathrm{BO}]_0$ gives $[\alpha] = -6$ dB and hence from Eq. (14.11)

$$[K] = [\alpha] + [B_{\mathrm{TR}}] - [B]$$
$$= -6 + 75.56 - 64.77$$
$$= 4.79 \text{ dB}$$

Hence

$$K = 10^{4.79/10} = \underline{\underline{3 \text{ (rounded down).}}}$$

If backoff was not required, the number of carriers which could be accommodated would be $B_{\mathrm{TR}}/B = 12$

14.7 TDMA

With TDMA, only one carrier uses the transponder at any one time, and therefore, intermodulation products, which result from the nonlinear amplification of multiple carriers, are absent. This leads to one of the most significant advantages of TDMA, which is that the TWT can be operated at maximum power output or saturation level.

Because the signal information is transmitted in bursts, TDMA is only suited to digital signals. Digital data can be assembled into burst format for transmission and reassembled from the received bursts through the use of digital buffer memories.

Figure 14.10 illustrates the basic TDMA concept, in which the stations transmit bursts in sequence. Burst synchronization is required, and in the system illustrated in Fig. 14.10, one station is assigned solely for the purpose of transmitting *reference bursts* to which the others can be synchronized. The time interval from the start of one reference burst to the next is termed a *frame*. A frame contains the reference burst R and the bursts from the other earth stations, these being shown as A, B, and C in Fig. 14.10.

Figure 14.11 illustrates the basic principles of burst transmission for a single channel. Overall, the transmission appears continuous because the input and output bit rates are continuous and equal. However, within the transmission channel, input bits are temporarily stored and transmitted in bursts. Since the time interval between bursts is the frame time T_F, the required buffer capacity is

$$M = R_b T_F \qquad (14.15)$$

The buffer memory fills up at the input bit rate R_b during the frame time interval. These M bits are transmitted as a burst in the next frame without any break in continuity of the input. The M bits are transmitted

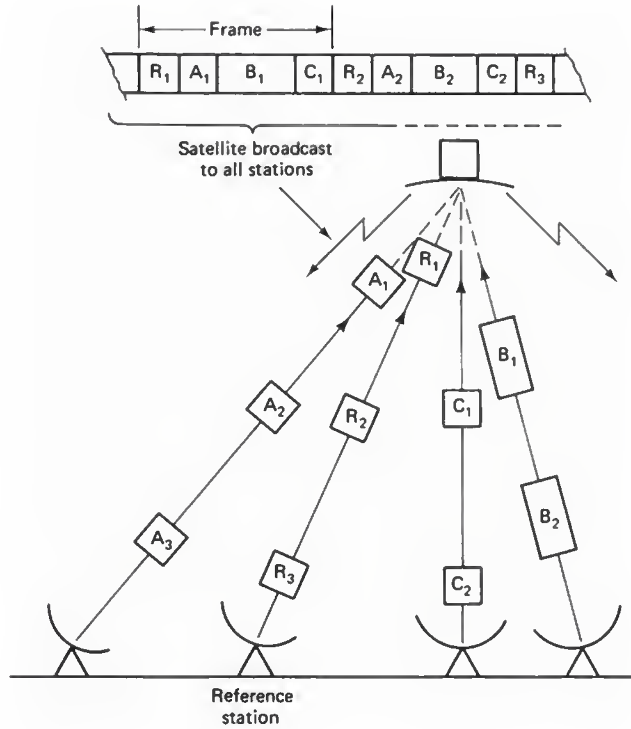


Figure 14.10 Time-division multiple access (TDMA) using a reference station for burst synchronization.

in the burst time T_B, and the *transmission rate*, which is equal to the burst bit rate, is

$$R_{\text{TDMA}} = \frac{M}{T_B} = R_b \frac{T_F}{T_B} \tag{14.16}$$

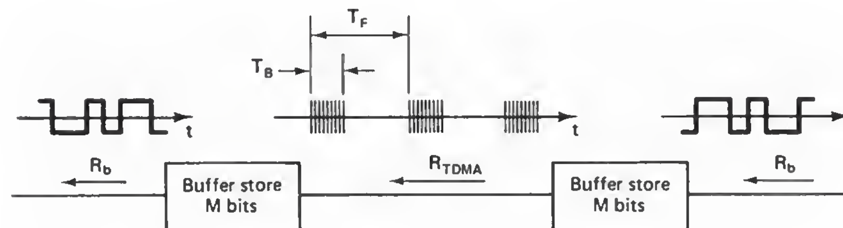


Figure 14.11 Burst-mode transmission linking two continuous-mode streams.

This is also referred to as the *burst rate*, but note that this means the instantaneous bit rate within a burst (not the number of bursts per second, which is simply equal to $1/T_B$). It will be seen that the *average* bit rate for the burst mode is simply M/T_F, which is equal to the input and output rates.

The frame time T_F will be seen to add to the overall propagation delay. For example, in the simple system illustrated in Fig. 14.11, even if the actual propagation delay between transmit and receive buffers is assumed to be zero, the receiving side would still have to wait a time T_F before receiving the first transmitted burst. In a geostationary satellite system, the actual propagation delay is a significant fraction of a second, and excessive delays from other causes must be avoided. This sets an upper limit to the frame time, although with current technology other factors restrict the frame time to well below this limit. The frame period is usually chosen to be a multiple of 125 μs, which is the standard sampling period used in *pulse-code modulation* (PCM) telephony systems, since this ensures that the PCM samples can be distributed across successive frames at the PCM sampling rate.

Figure 14.12 shows some of the basic units in a TDMA ground station, which for discussion purposes is labeled earth station *A*. Terrestrial links coming into earth station *A* carry digital traffic addressed to destination stations, labeled *B, C, X*. It is assumed that the bit rate is the same for the digital traffic on each terrestrial link. In the units labeled *terrestrial interface modules* (TIMs), the incoming continuous-bit-rate signals are converted into the intermittent-burst-rate mode. These individual burst-mode signals are *time-division multiplexed* in the time division multiplexer (MUX) so that the traffic for each destination station appears in its assigned time slot within a burst.

Certain time slots at the beginning of each burst are used to carry timing and synchronizing information. These time slots collectively are referred to as the *preamble*. The complete burst containing the preamble and the traffic data is used to phase modulate the *radiofrequency* (rf) carrier. Thus the composite burst which is transmitted at rf consists of a number of time slots, as shown in Fig. 14.13. These will be described in more detail shortly.

The received signal at an earth station consists of bursts from all transmitting stations arranged in the frame format shown in Fig. 14.13. The rf carrier is converted to *intermediate frequency* (IF), which is then demodulated. A separate preamble detector provides timing information for transmitter and receiver along with a carrier synchronizing signal for the phase demodulator, as described in the following section. In many systems, a station receives its own transmission along with the others in the frame, which can then be used for burst-timing purposes.

A reference burst is required at the beginning of each frame to provide timing information for the *acquisition* and *synchronization* of bursts

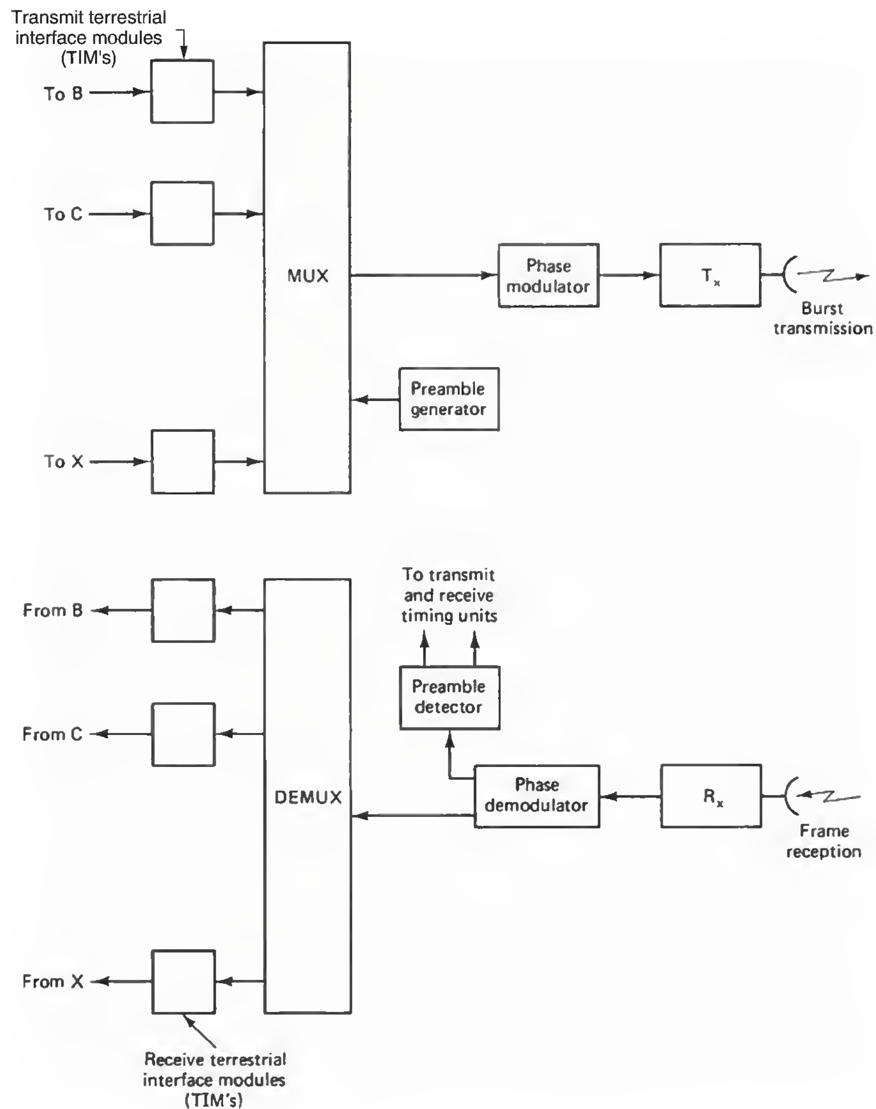


Figure 14.12 Some of the basic equipment blocks in a TDMA system.

(these functions are described further in Sec. 14.7.4). In the INTELSAT international network, at least two reference stations are used, one in the East and one in the West. These are designated *primary* reference stations, one of which is further selected as the *master primary*. Each primary station is duplicated by a *secondary* reference station, making four reference stations in all. The fact that all the reference stations are identical means that any one can become the master primary. All the system timing is derived from the high-stability clock in the master primary,

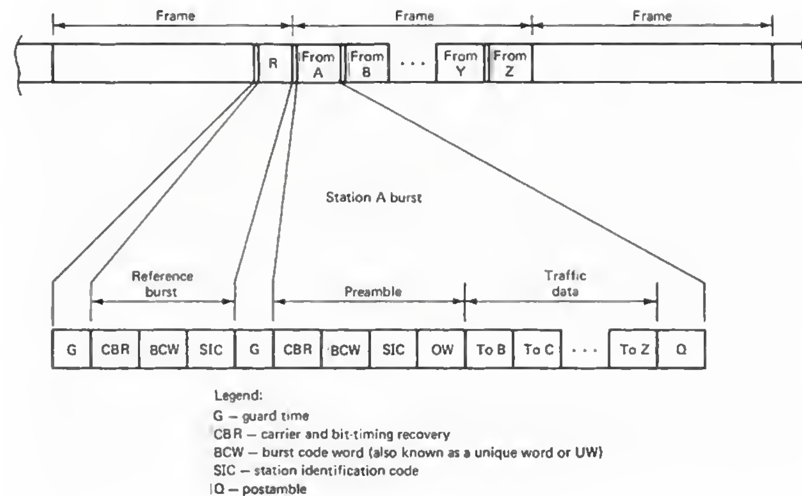


Figure 14.13 Frame and burst formats for a TDMA system.

which is accurate to 1 part in 10^{11} (Lewis, 1982). A clock on the satellite is locked to the master primary, and this acts as the clock for the other participating earth stations. The satellite clock will provide a constant frame time, but the participating earth stations must make corrections for variations in the satellite range, since the transmitted bursts from all the participating earth stations must reach the satellite in synchronism. Details of the timing requirements will be found in Spilker (1977).

In the INTELSAT system, two reference bursts are transmitted in each frame. The first reference burst, which marks the beginning of a frame, is transmitted by a master primary (or a primary) reference station and contains the timing information needed for the acquisition and synchronization of bursts. The second reference burst, which is transmitted by a secondary reference station, provides synchronization but not acquisition information. The secondary reference burst is ignored by the receiving earth stations unless the primary or master primary station fails.

14.7.1 Reference burst

The reference burst that marks the beginning of a frame is subdivided into time slots or channels used for various functions. These will differ in detail for different networks, but Fig. 14.13 shows some of the basic channels that are usually provided. These can be summarized as follows:

Guard time (G). A guard time is necessary between bursts to prevent the bursts from overlapping. The guard time will vary from burst to burst depending on the accuracy with which the various bursts can be positioned within each frame.

Carrier and bit-timing recovery (CBR). To perform coherent demodulation of the phase-modulated carrier, as described in Secs. 10.7 and 10.8, a coherent carrier signal must first be recovered from the burst. An unmodulated carrier wave is provided during the first part of the CBR time slot. This is used as a synchronizing signal for a local oscillator at the detector, which then produces an output coherent with the carrier wave. The carrier in the subsequent part of the CBR time slot is modulated by a known phase-change sequence which enables the bit timing to be recovered. Accurate bit timing is needed for the operation of the sample-and-hold function in the detector circuit (see Figs. 10.13 and 10.23). Carrier recovery is described in more detail in Sec. 14.7.3.

Burst code word (BCW). (Also known as a *unique word.*) This is a binary word, a copy of which is stored at each earth station. By comparing the incoming bits in a burst with the stored version of the BCW, the receiver can detect when a group of received bits matches the BCW, and this in turn provides an accurate time reference for the burst position in the frame. A known bit sequence is also carried in the BCW, which enables the phase ambiguity associated with coherent detection to be resolved.

Station identification code (SIC). This identifies the transmitting station.

Figure 14.14 shows the makeup of the reference bursts used in some of the INTELSAT networks. The numbers of symbols and the corresponding time intervals allocated to the various functions are shown. In addition to the channels already described, a *coordination and delay channel* or CDC (sometimes referred to as the *control and delay channel*) is provided. This channel carries the identification number of the earth station being addressed and various codes used in connection with the acquisition and synchronization of bursts at the addressed earth station. It is also necessary for an earth station to know the propagation time delay to the satellite to implement burst acquisition and synchronization. In the INTELSAT system, the propagation delay is computed from measurements made at the reference station and transmitted to the earth station in question through the coordination and delay channel.

The other channels in the INTELSAT reference burst are the following:

TTY: telegraph order-wire channel, used to provide telegraph communications between earth stations.

SC: service channel which carries various network protocol and alarm messages.

VOW: voice-order-wire channel used to provide voice communications between earth stations. Two VOW channels are provided.

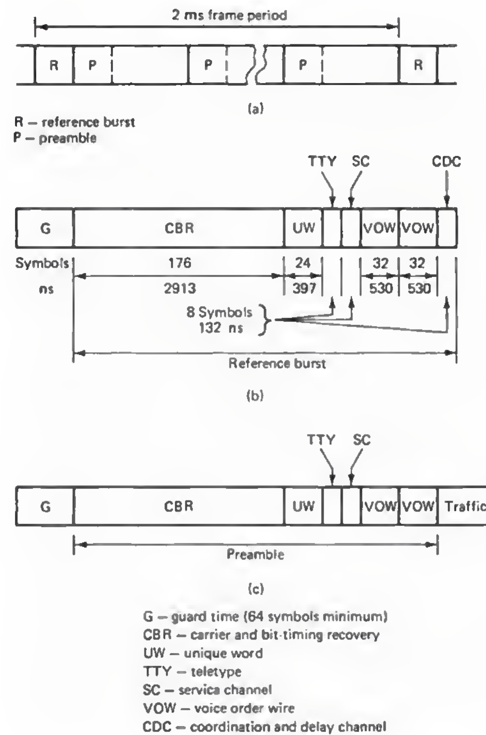


Figure 14.14 (*a*) Intelsat 2-ms frame; (*b*) composition of the reference burst *R*; (*c*) composition of the preamble *P*. (QPSK modulation is used, giving 2 bits per symbol. Approximate time intervals are shown.)

14.7.2 Preamble and postamble

The *preamble* is the initial portion of a traffic burst that carries information similar to that carried in the reference burst. In some systems the channel allocations in the reference bursts and the preambles are identical. No traffic is carried in the preamble. In Fig. 14.13, the only difference between the preamble and the reference burst is that the preamble provides an *orderwire* (OW) channel.

For the INTELSAT format shown in Fig. 14.14, the preamble differs from the reference burst in that it does not provide a CDC. Otherwise, the two are identical.

As with the reference bursts, the preamble provides a carrier and bit-timing recovery channel and also a burst-code-word channel for burst-timing purposes. The burst code word in the preamble of a traffic burst is different from the burst code word in the reference bursts, which enables the two types of bursts to be identified.

In certain phase detection systems, the phase detector must be allowed time to recover from one burst before the next burst is received by it. This is termed *decoder quenching*, and a time slot, referred to as a *postamble*, is allowed for this function. The postamble is shown as Q in Fig. 14.13. Many systems are designed to operate without a postamble.

14.7.3 Carrier recovery

A factor, which must be taken into account with TDMA is that the various bursts in a frame lack coherence so that carrier recovery must be repeated for each burst. This applies to the traffic as well as the reference bursts. Where the carrier recovery circuit employs a phase-locked loop such as shown in Fig. 10.20, a problem known as *hangup* can occur. This arises when the loop moves to an unstable region of its operating characteristic. The loop operation is such that it eventually returns to a stable operating point, but the time required to do this may be unacceptably long for burst-type signals.

One alternative method utilizes a narrowband tuned circuit filter to recover the carrier. An example of such a circuit for *quadrature phaseshift keying* (QPSK), taken from Miya (1981), is shown in Fig. 14.15. The QPSK signal, which has been downconverted to a standard IF of 140 MHz, is quadrupled in frequency to remove the modulation, as described in Sec. 10.7. The input frequency must be maintained at the resonant frequency of the tuned circuit, which requires some form of automatic frequency control. Because of the difficulties inherent in working with high frequencies, the output frequency of the quadrupler is downconverted from 560 to 40 MHz, and the AFC is applied to the *voltage controlled oscillator* (VCO) used to make the frequency conversion. The AFC circuit is a form of *phase-locked loop* (PLL), in which the phase difference between input and output of the single-tuned circuit is held at zero, which ensures that the 40-MHz input remains at the center of the tuned circuit response curve. Any deviation of the phase difference from zero generates a control voltage which is applied to the VCO in such a way as to bring the frequency back to the required value.

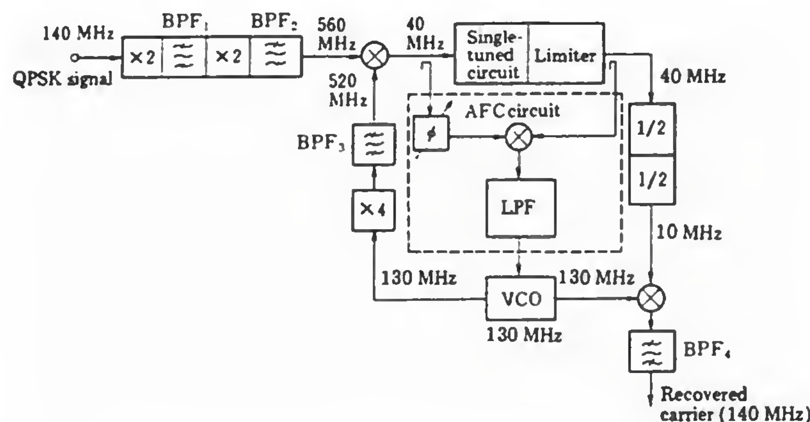


Figure 14.15 An example of carrier recovery circuit with a single-tuned circuit and AFC. (*Courtesy of Miya, 1981.*)

Interburst interference may be a problem with the tuned-circuit method because of the energy stored in the tuned circuit for any given burst. Avoidance of interburst interference requires careful design of the tuned circuit (Miya, 1981) and possibly the use of a postamble, as mentioned in the Sec. 14.7.2.

Other methods of carrier recovery are discussed in Gagliardi (1991).

14.7.4 Network synchronization

Network synchronization is required to ensure that all bursts arrive at the satellite in their correct time slots. As mentioned previously, timing markers are provided by the reference bursts, which are tied to a highly stable clock at the reference station and transmitted through the satellite link to the traffic stations. At any given traffic station, detection of the unique word (or burst code word) in the reference burst signals the *start of receiving frame* (SORF), the marker coinciding with the last bit in the unique word.

It would be desirable to have the highly stable clock located aboard the satellite because this would eliminate the variations in propagation delay arising from the uplink for the reference station, but this is not practical because of weight and space limitations. However, the reference bursts retransmitted from the satellite can be treated, for timing purposes, as if they originated from the satellite (Spilker, 1977).

The network operates what is termed a *burst time plan*, a copy of which is stored at each earth station. The burst time plan shows each earth station where the receive bursts intended for it are relative to the SORF marker. This is illustrated in Fig. 14.16. At earth station A the SORF marker is received after some propagation delay t_A, and

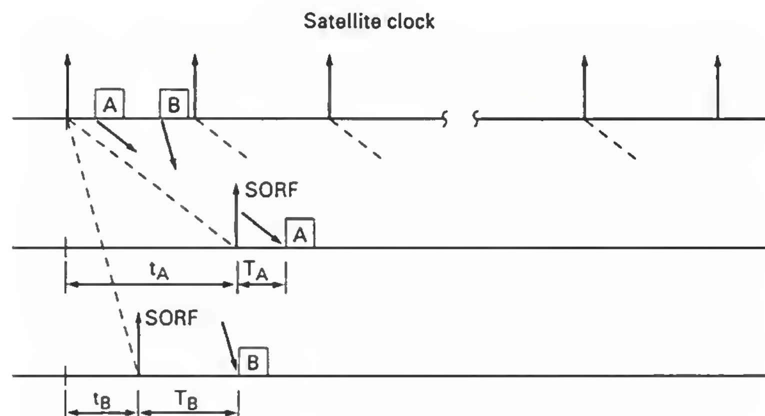


Figure 14.16 Start of receive frame (SORF) marker in a time burst plan.

the burst time plan tells station A that a burst intended for it follows at time T_A after the SORF marker received by it. Likewise, for station B, the propagation delay is t_B, and the received bursts start at T_B after the SORF markers received at station B. The propagation delays for each station will differ, but typically they are in the region of 120 ms each.

The burst time plan also shows a station when it must transmit its bursts in order to reach the satellite in the correct time slots. A major advantage of the TDMA mode of operation is that the burst time plan is essentially under software control so that changes in traffic patterns can be accommodated much more readily than is the case with FDMA, where modifications to hardware are required. Against this, implementation of the synchronization is a complicated process.

Corrections must be included for changes in propagation delay which result from the slowly varying position of the satellite (see Sec. 7.4). In general, the procedure for transmit timing control has two stages. First, there is the need for a station just entering, or reentering after a long delay, to acquire its correct slot position, this being referred to as *burst position acquisition*. Once the time slot has been acquired, the traffic station must maintain the correct position, this being known as *burst position synchronization*.

Open-loop timing control. This is the simplest method of transmit timing. A station transmits at a fixed interval following reception of the timing markers, according to the burst time plan, and sufficient guard time is allowed to absorb the variations in propagation delay. The burst position error can be large with this method, and longer guard times are necessary, which reduces frame efficiency (see Sec. 14.7.7). However, for frame times longer than about 45 ms, the loss of efficiency is less than 10 percent. In a modified version of the open-loop method known as *adaptive open-loop timing*, the range is computed at the traffic station from orbital data or from measurements, and the traffic earth station makes its own corrections in timing to allow for the variations in the range. It should be noted that with open-loop timing, no special acquisition procedure is required.

Loopback timing control. *Loopback* refers to the fact that an earth station receives its own transmission, from which it can determine range. It follows that the loopback method can only be used where the satellite transmits a global or regional beam encompassing all the earth stations in the network. A number of methods are available for the acquisition process (see, for example, Gagliardi, 1991), but basically, these all require some form of ranging to be carried out so that a close estimate of the slot position can be acquired. In one method, the traffic

station transmits a low-level burst consisting of the preamble only. The power level is 20 to 25 dB below the normal operating level (Ha, 1990) to prevent interference with other bursts, and the short burst is swept through the frame until it is observed to fall within the assigned time slot for the station. The short burst is then increased to full power, and fine adjustments in timing are made to bring it to the beginning of the time slot. Acquisition can take up to about 3 s in some cases. Following acquisition, the traffic data can be added, and synchronization can be maintained by continuously monitoring the position of the loopback transmission with reference to the SORF marker. The timing positions are reckoned from the last bit of the unique word in the preamble (as is also the case for the reference burst). The loopback method is also known as *direct closed-loop feedback.*

Feedback timing control. Where a traffic station lies outside the satellite beam containing its own transmission, loopback of the transmission does not of course occur, and some other method must be used for the station to receive ranging information. Where the synchronization information is transmitted back to an earth station from a distant station, this is termed *feedback closed-loop control*. The distant station may be a reference station, as in the INTELSAT network, or it may be another traffic station which is a designated *partner*. During the acquisition stage, the distant station can feed back information to guide the positioning of the short burst, and once the correct time slot is acquired, the necessary synchronizing information can be fed back on a continuous basis.

Figure 14.17 illustrates the feedback closed-loop control method for two earth stations A and B. The SORF marker is used as a reference point for the burst transmissions. However, the reference point which denotes the *start of transmit frame* (SOTF) has to be delayed by a certain amount, shown as D_A for earth station A and D_B for earth-station B. This is necessary so that the SOTF reference points for each earth station coincide at the satellite transponder, and the traffic bursts, which are transmitted at their designated times after the SOTF, arrive in their correct relative positions at the transponder, as shown in Fig. 14.17. The total time delay between any given satellite clock pulse and the corresponding SOTF is a constant, shown as C in Fig. 14.17. C is equal to $2t_A + D_A$ for station A and $2t_B + D_B$ for station B. In general, for earth station i, the delay D_i is determined by

$$2t_i + D_i = C \tag{14.17}$$

In the INTELSAT network, $C = 288$ ms.

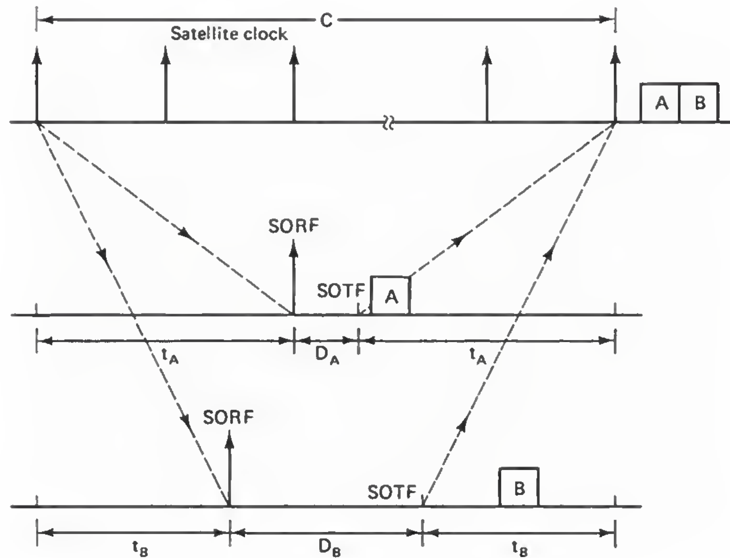


Figure 14.17 Timing relationships in a TDMA system. SORF, start of receive frame; SOTF, start of transmit frame.

For a truly geostationary satellite, the propagation delay t_i would be constant. However, as shown in Sec. 7.4, station-keeping maneuvers are required to keep a geostationary satellite at its assigned orbital position, and hence this position can be held only within certain tolerances. For example, in the INTELSAT network, the variation in satellite position can lead to a variation of up to ±0.55 ms in the propagation delay (INTELSAT, 1980). In order to minimize the guard time needed between bursts, this variation in propagation delay must be taken into account in determining the delay, D_i, required at each traffic station. In the INTELSAT network, the D_i numbers are updated every 512 frames, which is a period of 1.024 s, based on measurements and calculations of the propagation delay times made at the reference station. The D_i numbers are transmitted to the earth stations through the CDC channel in the reference bursts. (It should be noted that the open-loop synchronization described previously amounts to using a constant D_i value.)

The use of traffic burst preambles along with reference bursts to achieve synchronization is the most common method, but at least one other method, not requiring preambles, has been proposed by Nuspl and de Buda (1974). It also should be noted that there are certain types of "packet satellite networks," for example, the basic Aloha system (Rosner, 1982), which are closely related to TDMA, in which synchronization is not used.

14.7.5 Unique word detection

The *unique word* (UW) or *burst code word* (BCW) is used to establish burst timing in TDMA. Figure 14.18 shows the basic arrangement for detecting the UW. The received bit stream is passed through a shift register which forms part of a correlator. As the bit stream moves through the register, the sequence is continuously compared with a stored version of the UW. When correlation is achieved, indicated by a high output from the threshold detector, the last bit of the UW provides the reference point for timing purposes. It is important therefore to know the probability of error in detecting the UW. Two possibilities have to be considered. One, termed the *miss probability*, is the probability of the correlation detector failing to detect the UW even though it is present in the bit stream. The other, termed the *probability of false alarm*, is the probability that the correlation detector misreads a sequence as the UW. Both of these will be examined in turn.

Miss probability. Let E represent the maximum number of errors allowed in the UW of length N bits, and let I represent the actual number of errors in the UW as received. The following conditions apply:

When $I \leq E$, the detected sequence is declared to be the UW.

When $I > E$, the detected sequence N is declared not to be the UW; that is, the unique word is missed.

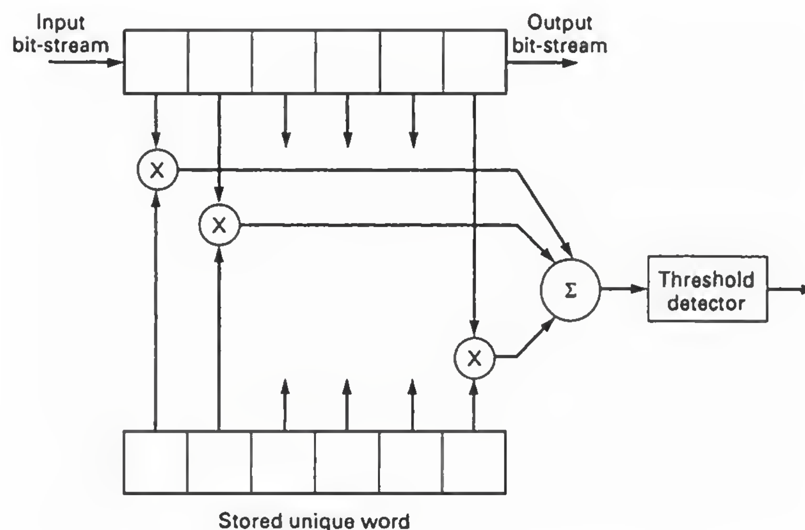


Figure 14.18 Basic arrangement for detection of the unique word (UW).

Let p represent the average probability of error in transmission [the *bit error rate* (BER)]. The probability of receiving a sequence N containing I errors in any one particular arrangement is

$$p_I = p^I(1 - p)^{N-I} \tag{14.18}$$

The number of combinations of N things taken I at a time, usually written as ${}_NC_I$, is given by

$${}_NC_I = \frac{N!}{I!(N - I)!} \tag{14.19}$$

The probability of receiving a sequence of N bits containing I errors is therefore

$$P_I = {}_NC_I \, p_I \tag{14.20}$$

Now since the UW is just such a sequence, Eq. (14.20) gives the probability of a UW containing I errors. The condition for a miss occurring is that $I > E$, and therefore, the miss probability is

$$P_{\text{miss}} = \sum_{I=E+1}^{N} P_I \tag{14.21}$$

Written out in full, this is

$$P_{\text{miss}} = \sum_{I=E+1}^{N} \frac{N!}{I!(N - I)!} p^I(1 - p)^{N-I} \tag{14.22}$$

Equation (14.22) gives the *average* probability of missing the UW even though it is present in the shift register of the correlator. Note that because this is an average probability, it is not necessary to know any specific value of I.

Example 14.2 Determine the miss probability for the following values: $N = 40$, $E = 5$, $p = 10^{-3}$

Solution

$$P_{\text{miss}} = \sum_{I=6}^{40} \frac{40!}{I!(40 - I)!} 10^{-3I}(1 - 10^{-3})^{40-I}$$
$$= 3.7 \times 10^{-12}$$

False detection probability. Consider now a sequence of N which is not the UW but which would be interpreted as the UW even if it differs

from it in some number of bit positions E, and let I represent the number of bit positions by which the random sequence actually does differ from the UW. Thus E represents the number of acceptable "bit errors" considered from the point of view of the UW, although they may not be errors in the message they represent. Likewise, I represents the actual number of "bit errors" considered from the point of view of the UW, although they may not be errors in the message they represent. As before, the number of combinations of N things taken I at a time is given by Eq. (14.19), and hence the number of words acceptable as the UW is

$$W = \sum_{I=0}^{E} {}_NC_I \tag{14.23}$$

The number of words which can be formed from a random sequence of N bits is 2^N, and on the assumption that all such words are equiprobable, the probability of receiving any one particular word is 2^{-N}. Hence the probability of a false detection is

$$P_F = 2^{-N}W \tag{14.24}$$

Written out in full, this is

$$P_F = 2^{-N}\sum_{I=0}^{E}\frac{N!}{I!(N-I)!} \tag{14.25}$$

Again it will be noticed that because this is an average probability, it is not necessary to know a specific value of I. Also, in this case, the BER does not enter into the calculation.

Example 14.3 Determine the probability of false detection for the following values: $N = 40$, $E = 5$

Solution

$$P_F = 2^{-40}\sum_{I=0}^{5}\frac{40!}{I!(40-I)!}$$

$$= \underline{\underline{6.9 \times 10^{-7}}}$$

From Examples 14.2 and 14.3 it is seen that the probability of a false detection is much higher than the probability of a miss, and this is true in general. In practice, once frame synchronization has been established, a time window can be formed around the expected time of arrival for the UW such that the correlation detector is only in operation for the window period. This greatly reduces the probability of false detection.

14.7.6 Traffic data

The traffic data immediately follow the preamble in a burst. As shown in Fig. 14.13, the traffic data subburst is further subdivided into time slots addressed to the individual destination stations. Any given destination station selects only the data in the time slots intended for that station. As with FDMA networks, TDMA networks can be operated with both preassigned and demand assigned channels, and examples of both types will be given shortly. The greater the fraction of frame time that is given over to traffic, the higher is the efficiency. The concept of *frame efficiency* is discussed in the following section.

14.7.7 Frame efficiency and channel capacity

The frame efficiency is a measure of the fraction of frame time used for the transmission of traffic. *Frame efficiency* may be defined as

$$\text{Frame efficiency} = \eta_F = \frac{\text{traffic bits}}{\text{total bits}} \tag{14.26}$$

Alternatively, this can be written as

$$\eta_F = 1 - \frac{\text{overhead bits}}{\text{total bits}} \tag{14.27}$$

In these equations, bits per frame are implied. The overhead bits consist of the sum of the preamble, the postamble, the guard intervals, and the reference-burst bits per frame. The equations may be stated in terms of symbols rather than bits, or the actual times may be used.

For a fixed overhead, Eq. (14.27) shows that a longer frame, or greater number of total bits, results in higher efficiency. However, longer frames require larger buffer memories and also add to the propagation delay. Synchronization also may be made more difficult, keeping in mind that the satellite position is varying with time. It is clear that a lower overhead also leads to higher efficiency, but again, reducing synchronizing and guard times may result in more complex equipment being required.

Example 14.4 Calculate the frame efficiency for an INTELSAT frame given the following information:

Total frame length = 120,832 symbols

Traffic bursts per frame = 14

Reference bursts per frame = 2

Guard interval = 103 symbols

Solution From Fig. 14.14, the preamble symbols add up to

$$P = 176 + 24 + 8 + 8 + 32 + 32 = 280$$

With addition of the CDC channel, the reference channel symbols add up to

$$R = 280 + 8 = 288$$

Therefore, the overhead symbols are

$$\text{OH} = 2 \times (103 + 288) + 14 \times (103 + 280) = 6144 \text{ symbols}$$

Therefore, from Eq. (14.27),

$$\eta_F = 1 - \frac{6144}{120832} = \underline{\underline{0.949}}$$

The voice-channel capacity of a frame, which is also the voice-channel capacity of the transponder being accessed by the frame, can be found from a knowledge of the frame efficiency and the bit rates. Let R_b be the bit rate of a voice channel, and let there be a total of n voice channels shared between all the earth stations accessing the transponder. The total incoming *traffic* bit rate to a frame is nR_b. The traffic bit rate of the frame is $\eta_F R_{\text{TDMA}}$, and therefore

$$nR_b = \eta_F R_{\text{TDMA}}$$

or

$$n = \frac{\eta_F R_{\text{TDMA}}}{R_b} \tag{14.28}$$

Example 14.5 Calculate the voice-channel capacity for the INTELSAT frame in Example 14.2, given that the voice-channel bit rate is 64 kb/s and that QPSK modulation is used. The frame period is 2 ms.

Solution The number of symbols per frame is 120,832, and the frame period is 2 ms. Therefore, the symbol rate is 120,832/2 ms = 60.416 megasymbols/s. QPSK modulation utilizes 2 bits per symbol, and therefore, the transmission rate is $R_{\text{TDMA}} = 60.416 \times 2 = 120.832$ Mb/s.

Using Eq. (14.28) and the efficiency as calculated in Example 14.4,

$$n = 0.949 \times 120.832 \times \frac{10^3}{64} = \underline{\underline{1792}}$$

14.7.8 Preassigned TDMA

An example of a preassigned TDMA network is the CSC for the Spade network described in Sec. 14.5. The frame and burst formats are shown in Fig. 14.19. The CSC can accommodate up to 49 earth stations in the network plus one reference station, making a maximum of 50 bursts in a frame.

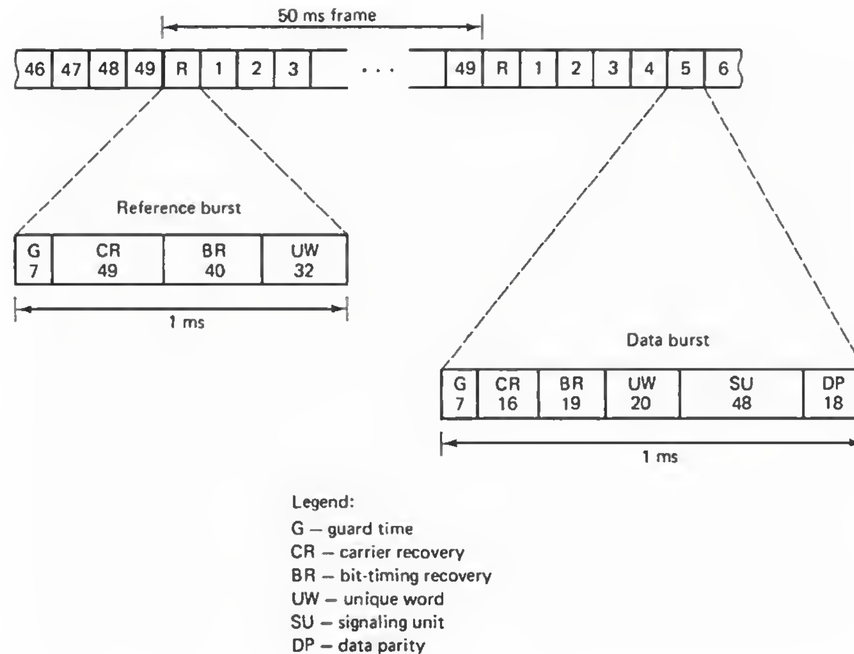


Figure 14.19 Frame and bit formats for the common signaling channel (CSC) used with the Spade system. (*Data from Miya, 1981.*)

All the bursts are of equal length. Each burst contains 128 bits and occupies a 1-ms time slot. Thus the bit rate is 128 kb/s. As discussed in Sec. 14.5, the frequency bandwidth required for the CSC is 160 kHz.

The *signaling unit* (SU) shown in Fig. 14.19 is that section of the data burst which is used to update the other stations on the status of the frequencies available for the SCPC calls. It also carries the signaling information, as described in Sec. 14.5.

Another example of a preassigned TDMA frame format is the INTELSAT frame shown in simplified form in Fig. 14.20. In the INTELSAT system, preassigned and demand-assigned voice channels are carried together, but for clarity, only a preassigned traffic burst is shown. The traffic burst is subdivided into time slots, termed *satellite channels* in the INTELSAT terminology, and there can be up to 128 of these in a traffic burst. Each satellite channel is further subdivided into 16 time slots termed *terrestrial channels*, each terrestrial channel carrying one PCM sample of an analog telephone signal. QPSK modulation is used, and therefore, there are 2 bits per symbol as shown. Thus each terrestrial channel carries 4 symbols (or 8 bits). Each satellite channel carries $4 \times 16 = 64$ symbols, and at its maximum of 128 satellite channels, the traffic burst carries 8192 symbols.

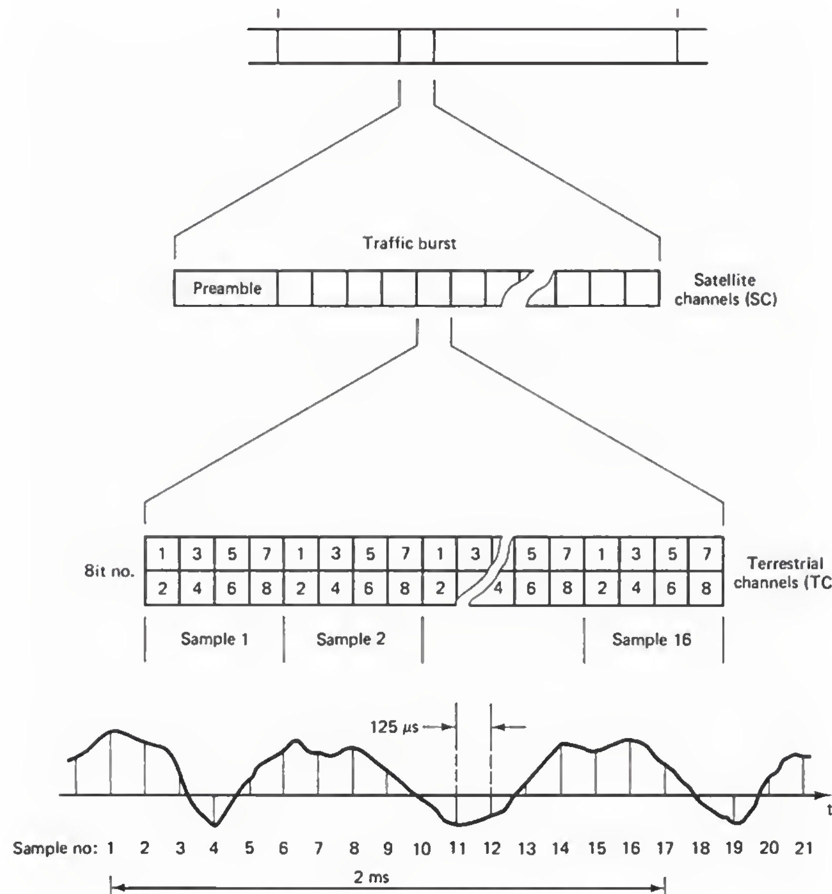


Figure 14.20 Preassigned TDMA frame in the Intelsat system.

As discussed in Sec. 10.3, the PCM sampling rate is 8 kHz, and with 8 bits per sample, the PCM bit rate is 64 kb/s. Each satellite channel can accommodate this bit rate. Where input data at a higher rate must be transmitted, multiple satellite channels are used. The maximum input data rate which can be handled is 128(SC) × 64 kb/s = 8.192 Mb/s.

The INTELSAT frame is 120,832 symbols or 241,664 bits long. The frame period is 2 ms, and therefore, the burst bit rate is 120.832 Mb/s.

As mentioned previously, preassigned and demand-assigned voice channels can be accommodated together in the INTELSAT frame format. The demand-assigned channels utilize a technique known as *digital speech interpolation* (DSI), which is described in the following section.

The preassigned channels are referred to as *digital noninterpolated* (DNI) channels.

14.7.9 Demand-assigned TDMA

With TDMA, the burst and subburst assignments are under software control, compared with hardware control of the carrier frequency assignments in FDMA. Consequently, compared with FDMA networks, TDMA networks have more flexibility in reassigning channels, and the changes can be made more quickly and easily.

A number of methods are available for providing traffic flexibility with TDMA. The burst length assigned to a station may be varied as the traffic demand varies. A central control station may be employed by the network to control the assignment of burst lengths to each participating station. Alternatively, each station may determine its own burst-length requirements and assign these in accordance with a prearranged network discipline.

As an alternative to burst-length variation, the burst length may be kept constant and the number of bursts per frame used by a given station varied as demand requires. In one proposed system (CCIR Report 708, 1982), the frame length is fixed at 13.5 ms. The basic burst time slot is 62.5 μs, and stations in the network transmit information bursts varying in discrete steps over the range 0.5 ms (8 basic bursts) to 4.5 ms (72 basic bursts) per frame. Demand assignment for speech channels takes advantage of the intermittent nature of speech, as described in the following section.

14.7.10 Speech interpolation and prediction

Because of the intermittent nature of speech, a speech transmission channel lies inactive for a considerable fraction of the time it is in use. A number of factors contribute to this. The talk-listen nature of a normal two-way telephone conversation means that transmission in any one direction occurs only about 50 percent of the time. In addition, the pauses between words and phrases may further decrease this to about 33 percent. If further allowance is made for "end party" delays such as the time required for a party to answer a call, the average fraction of the total connect time may drop to as low as 25 percent. The fraction of time a transmission channel is active is known as the *telephone load activity factor*, and for system design studies, the value of 0.25 is recommended by *Comité Consutatif Internationale Télégraphique et Téléphonique* (CCITT), although higher values are also used (Pratt and Bostian, 1986). The point is that for a significant fraction of the time the channel is available for other transmissions, and advantage is taken of this in the form of demand assignment known as *digital speech interpolation.*

Digital speech interpolation may be implemented in one of two ways, these being digital *time assignment speech interpolation* (digital TASI) and *speech predictive encoded communications* (SPEC).

Digital TASI. The traffic-burst format for an INTELSAT burst carrying demand-assigned channels and preassigned channels is shown in Fig. 14.21. As mentioned previously, the demand-assigned channels utilize digital TASI, or what is referred to in the INTELSAT nomenclature as DSI, for *digital speech interpolation*. These are shown by the block labeled "interpolated" in Fig. 14.21. The first satellite channel (channel 0) in this block is an assignment channel, labeled *DSI-AC*. No traffic is carried in the assignment channel; it is used to transmit channel assignment information as will be described shortly.

Figure 14.22 shows in outline the DSI system. Basically, the system allows N terrestrial channels to be carried by M satellite channels, where $N > M$. For example, in the INTELSAT arrangement, $N = 240$ and $M = 127$.

On each incoming terrestrial channel, a speech detector senses when speech is present, the intermittent speech signals being referred to as *speech spurts*. A speech spurt lasts on average about 1.5 s (Miya, 1981). A control signal from the speech detector is sent to the channel assignment unit, which searches for an empty TDMA buffer. Assuming that one is found, the terrestrial channel is assigned to this satellite channel, and the speech spurt is stored in the buffer, ready for transmission in the DSI subburst. A delay is inserted in the speech circuit, as shown in Fig. 14.22, to allow some time for the assignment process to

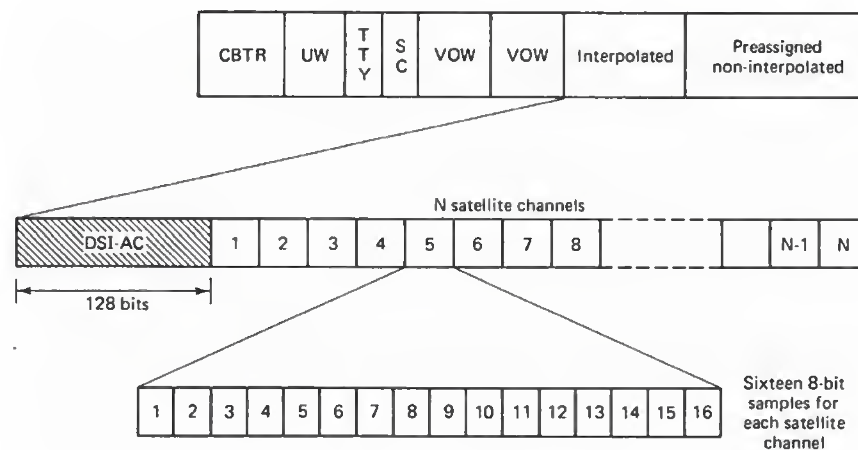


Figure 14.21 Intelsat traffic burst structure. (*Courtesy of Intelsat, 1983. With permission.*)

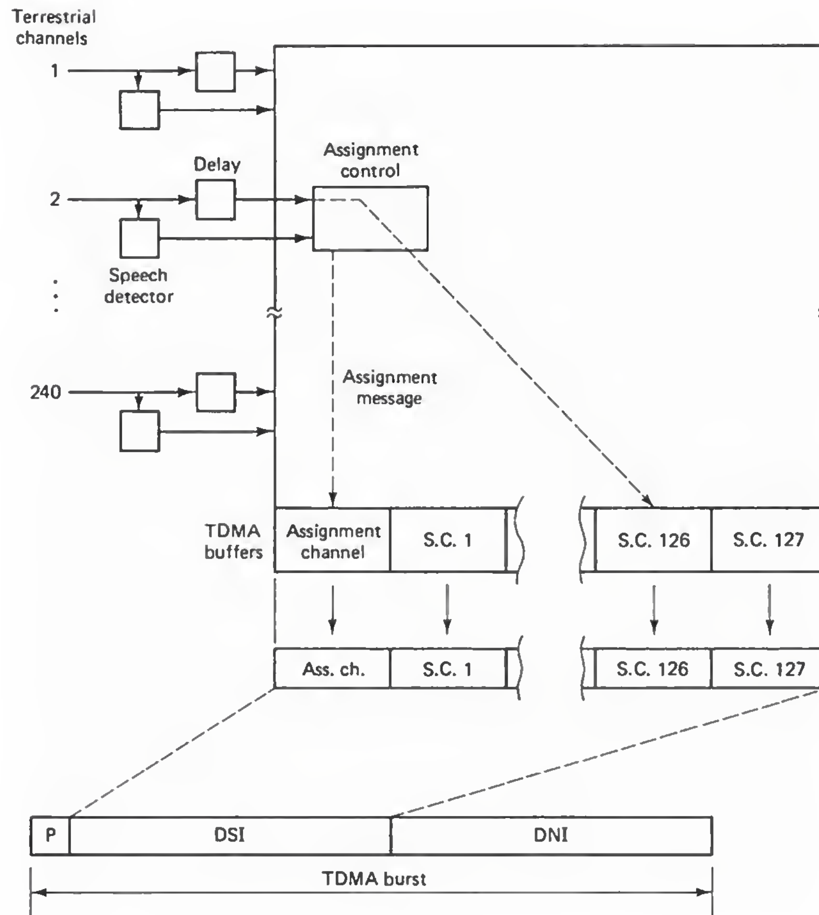


Figure 14.22 Digital speech interpolation. DSI = digital speech interpolation; DNI = digital noninterpolation.

be completed. However, this delay cannot exactly compensate for the assignment delay, and the initial part of the speech spurt may be lost. This is termed a *connect clip*.

In the INTELSAT system an intermediate step occurs where the terrestrial channels are renamed *international channels* before being assigned to a satellite channel (Pratt and Bostian, 1986). For clarity, this step is not shown in Fig. 14.22.

At the same time as an assignment is made, an assignment message is stored in the assignment channel buffer, which informs the receive stations which terrestrial channel is assigned to which satellite channel. Once an assignment is made, it is not interrupted, even during pauses between spurts, unless the pause times are required for another DSI

channel. This reduces the amount of information needed to be transmitted over the assignment channel.

At the receive side, the traffic messages are stored in their respective satellite-channel buffers. The assignment information ensures that the correct buffer is read out to the corresponding terrestrial channel during its sampling time slot. During speech pauses when the channel has been reassigned, a low-level noise signal is introduced at the receiver to simulate a continuous connection.

It has been assumed that a free satellite channel will be found for any incoming speech spurt, but of course there is a finite probability that all channels will be occupied and the speech spurt lost. Losing a speech spurt in this manner is referred to as *freeze-out*, and the freeze-out fraction is the ratio of the time the speech is lost to the average spurt duration. It is found that a design objective of 0.5 percent for a freeze-out fraction is satisfactory in practice. This means that the probability of a freeze-out occurring is 0.005.

Another source of signal mutilation is the *connect clip* mentioned earlier. Again, it is found in practice that clips longer than about 50 ms are very annoying to the listener. An acceptable design objective is to limit the fraction of clips which are equal to or greater than 50 ms to a maximum of 2 percent of the total clips. In other words, the probability of encountering a clip that exceeds 50 ms is 0.02.

The *DSI gain* is the ratio of the number of terrestrial channels to number of satellite channels, or N/M. The DSI gain depends on the number of satellite channels provided as well as the design objectives stated earlier. Typically, DSI gains somewhat greater than 2 can be achieved in practice.

Speech predictive encoded communications (SPEC). The block diagram for the SPEC system is shown in Fig. 14.23 (Sciulli and Campanella, 1973). In this method, the incoming speech signals are converted to a PCM multiplexed signal using 8 bits per sample quantization. With 64 input lines and sampling at 125 μs, the output bit rate from the multiplexer is $8 \times 64/125 = 4.096$ Mb/s.

The digital voice switch following the PCM multiplexer is time-shared between the input signals. It is voice-activated to prevent transmission of noise during silent intervals. When the zero-order predictor receives a new sample, it compares it with the previous sample for that voice channel, which it has stored, and transmits the new sample only if it differs from the preceding one by a predetermined amount. These new samples are labeled *unpredictable PCM samples* in Fig. 14.23*a*.

For the 64 channels a 64-bit assignment word is also sent. A logic 1 in the channel for the assignment word means that a new sample was sent

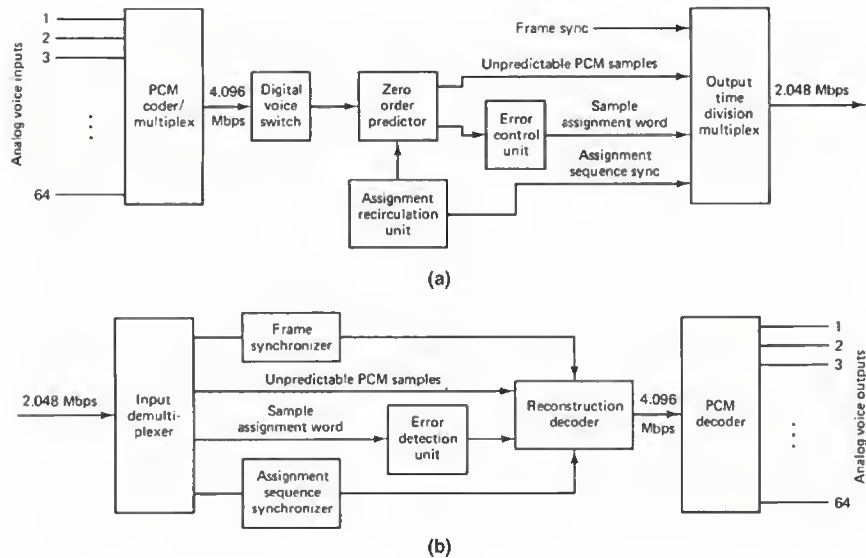


Figure 14.23 (*a*) SPEC transmitter; (*b*) SPEC receiver. (*Courtesy of Sciulli and Campanella, 1973. Copyright 1973, IEEE.*)

for that channel, and a logic 0 means that the sample was unchanged. At the receiver, the sample assignment word either directs the new (unpredictable) sample into the correct channel slot, or it results in the previous sample being regenerated in the reconstruction decoder. The output from this is a 4.096-Mb/s PCM multiplexed signal which is demultiplexed in the PCM decoder.

By removing the redundant speech samples and silent periods from the transmission link, a doubling in channel capacity is achieved. As shown in Fig. 14.23, the transmission is at 2.048 Mb/s for an input-output rate of 4.096 Mb/s.

An advantage of the SPEC method over the DSI method is that freeze-out does not occur during overload conditions. During overload, sample values which should change may not. This effectively leads to a coarser quantization and therefore an increase in quantization noise. This is subjectively more tolerable than freeze-out.

14.7.11 Downlink analysis for digital transmission

As mentioned in Sec. 14.6, the transponder power output and bandwidth both impose limits on the system transmission capacity. With TDMA, TWT backoff is not generally required, which allows the transponder to operate at saturation. One drawback arising from this is that the uplink

station must be capable of saturating the transponder, which means that even a low-traffic-capacity station requires comparatively large power output compared with what would be required for FDMA. This point is considered further in Sec. 14.7.12.

As with the FDM/FDMA system analysis, it will be assumed that the overall carrier-to-noise ratio is essentially equal to the downlink carrier-to-noise ratio. With a power-limited system this C/N ratio is one of the factors that determines the maximum digital rate, as shown by Eq. (10.24). Equation (10.24) can be rewritten as

$$[R_b] = \left[\frac{C}{N_0}\right] - \left[\frac{E_b}{N_0}\right] \tag{14.29}$$

The $[E_b/N_0]$ ratio is determined by the required BER, as shown in Fig. 10.17, and described in Sec. 10.6.4. For example, for a BER of 10^{-5} an $[E_b/N_0]$ of 9.6 dB is required. If the rate R_b is specified, then the $[C/N_0]$ ratio is determined, as shown by Eq. (14.29), and this value is used in the link budget calculations as required by Eq. (12.53). Alternatively, if the $[C/N_0]$ ratio is fixed by the link budget parameters as given by Eq. (12.53), the bit rate is then determined by Eq. (14.29).

The bit rate is also constrained by the IF bandwidth. As shown in Sec. 10.6.3, the ratio of bit rate to IF bandwidth is given by

$$\frac{R_b}{B_{\mathrm{IF}}} = \frac{m}{1 + \rho}$$

where $m = 1$ for *binary phase-shift keying* (BPSK) and $m = 2$ for QPSK and ρ is the rolloff factor. The value of 0.2 is commonly used for the rolloff factor, and therefore, the bit rate for a given bandwidth becomes

$$R_b = \frac{mB_{\mathrm{IF}}}{1.2} \tag{14.30}$$

Example 14.6 Using Eq. (12.53), a downlink $[C/N_0]$ of 87.3 dBHz is calculated for a TDMA circuit that uses QPSK modulation. A BER of 10^{-5} is required. Calculate the maximum transmission rate. Calculate also the IF bandwidth required assuming a rolloff factor of 0.2.

Solution From Fig. 10.18 which is applicable for QPSK and BPSK. $[E_b/N_0] = 9.65$ dB for a BER of 10^{-5}. Hence

$$[R_b] = 87.3 - 9.65 = 77.65 \text{ dBb/s}$$

This is equal to 58.21 Mb/s.

For QPSK $m = 2$ and using Eq. (14.30), we have

$$B_{IF} = 58.21 \times 1.2/2 = \underline{\underline{34.9 \text{ MHz}}}$$

From Example 14.6 it will be seen that if the satellite transponder has a bandwidth of 36 MHz, and an EIRP that results in a $[C/N_0]$ of 87.3 dBHz at the receiving ground station, the system is rear optimum in that the bandwidth is almost fully occupied.

14.7.12 Comparison of uplink power requirements for FDMA and TDMA

With FDMA, the modulated carriers at the input to the satellite are retransmitted from the satellite as a combined frequency-division-multiplexed signal. Each carrier retains its modulation, which may be analog or digital. For this comparison, digital modulation will be assumed. The modulation bit rate for each carrier is equal to the input bit rate [adjusted as necessary for *forward error correction* (FEC)]. The situation is illustrated in Fig. 14.24*a*, where for simplicity, the input bit rate R_b is assumed to be the same for each earth station. The [EIRP] is also assumed to be the same for each earth station.

With TDMA, the uplink bursts that are displaced in time from one another are retransmitted from the satellite as a combined time-division-multiplexed signal. The uplink bit rate is equal to the downlink bit rate in this case, as illustrated in Fig. 14.24*b*. As described in Sec. 14.7, compression buffers are needed in order to convert the input bit rate R_b to the transmitted bit rate R_{TDMA}.

Because the TDMA earth stations have to transmit at a higher bit rate compared with FDMA, a higher [EIRP] is required, as can be deduced from Eq. (10.24). Equation (10.24) states that

$$\left[\frac{C}{N_0}\right] = \left[\frac{E_b}{N_0}\right] + [R]$$

where $[R]$ is equal to $[R_b]$ for an FDMA uplink and $[R_{TDMA}]$ for a TDMA uplink.

For a given BER the $[E_b/N_0]$ ratio is fixed as shown by Fig. 10.17. Hence, assuming that $[E_b/N_0]$ is the same for the TDMA and the FDMA uplinks, an increase in $[R]$ requires a corresponding increase in $[C/N_0]$. Assuming that the TDMA and FDMA uplinks operate with the same [LOSSES] and satellite $[G/T]$, Eq. (12.39) shows that the increase in $[C/N_0]$ can be achieved only through an increase in the earth station [EIRP], and therefore

$$[\text{EIRP}]_{TDMA} - [\text{EIRP}]_{FDMA} = [R_{TDMA}] - [R_b] \qquad (14.31)$$

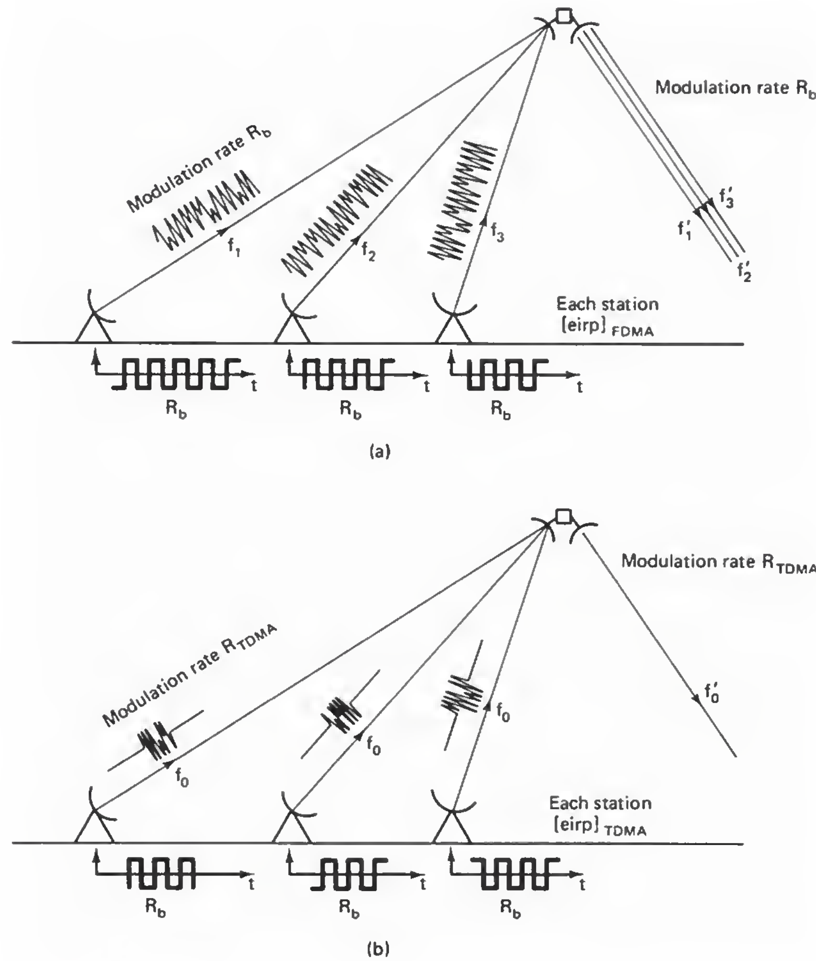


Figure 14.24 (*a*) FDMA network; (*b*) TDMA network.

For the same earth-station antenna gain in each case, the decibel increase in earth station transmit power for TDMA compared with FDMA is

$$[P]_{\mathrm{TDMA}} - [P]_{\mathrm{FDMA}} = [R_{\mathrm{TDMA}}] - [R_b] \tag{14.32}$$

Example 14.7 A 14-GHz uplink operates with transmission losses and margins totaling 212 dB and a satellite $[G/T] = 10$ dB/K. The required uplink $[E_b/N_0]$ is 12 dB. (*a*) Assuming FDMA operation and an earth-station uplink antenna gain of 46 dB, calculate the earth-station transmitter power needed for transmission of a T1 baseband signal. (*b*) If the downlink transmission

rate is fixed at 74 dBb/s, calculate the uplink power increase required for TDMA operation.

Solution (*a*) From Sec. 10.4 the T1 bit rate is 1.544 Mb/s or $[R] = 62$ dBb/s.

Using the $[E_b/N_0] = 12$-dB value specified, Eq. (10.24) gives

$$\left[\frac{C}{N_0}\right] = 12 + 62 = 74 \text{ dBHz}$$

From Eq. (12.39),

$$\begin{aligned}[\text{EIRP}] &= \left[\frac{C}{N_0}\right] - \left[\frac{G}{T}\right] + [\text{LOSSES}] - 228.6 \\ &= 74 - 10 + 212 - 228.6 \\ &= 47.4 \text{ dBW}\end{aligned}$$

Hence the transmitter power required is

$$[P] = 47.4 - 46 = \underline{\underline{1.4 \text{ dBW} \quad \text{or} \quad 1.38 \text{ W}}}$$

(*b*) With TDMA operation the rate increase is $74 - 62 = 12$ dB. All other factors being equal, the earth station [EIRP] must be increased by this amount, and hence

$$[P] = 1.4 + 12 = \underline{\underline{13.4 \text{ dBW} \quad \text{or} \quad 21.9 \text{ W}}}$$

For small satellite business systems it is desirable to be able to operate with relatively small earth stations, which suggests that FDMA should be the mode of operation. On the other hand, TDMA permits more efficient use of the satellite transponder by eliminating the need for backoff. This suggests that it might be worthwhile to operate a hybrid system in which FDMA is the uplink mode of operation, with the individual signals converted to a time-division-multiplexed format in the transponder before being amplified by the TWTA. This would allow the transponder to be operated at saturation as in TDMA. Such a hybrid mode of operation would require the use of a signal-processing transponder as discussed in the following section.

14.8 On-Board Signal Processing for FDMA/TDM Operation

As seen in the preceding section, for small earth stations carrying digital signals at relatively low data rates, there is an advantage to be gained in terms of earth station power requirements by using FDMA. On the other hand, TDMA signals make more efficient use of the transponder because back-off is not required.

Market studies show that what is termed *customer premises services* (CPS) will make up a significant portion of the satellite demand over the decade 1990–2000 (Stevenson et al., 1984). Multiplexed digital transmission will be used, most likely at the T1 rate. This bit rate provides for most of the popular services, such as voice, data, and videoconferencing, but specifically excludes standard television signals. Customer premises services is an ideal candidate for the FDMA/TDM mode of operation mentioned in the preceding section. To operate in this mode requires the use of *signal-processing transponders*, in which the FDMA uplink signals are converted to the TDM format for retransmission on the downlink. It also should be noted that the use of signal processing transponders "decouples" the uplink from the downlink. This is important because it allows the performance of each link to be optimized independently of the other.

A number of signal-processing methods have been proposed. One conventional approach is illustrated in the simplified block schematic of Fig. 14.25*a*. Here the individual uplink carriers at the satellite are selected by frequency filters and detected in the normal manner. The baseband signals are then combined in the baseband processor, where they are converted to a time-division-multiplexed format for remodulation onto a downlink carrier. More than one downlink carrier may be provided, but only one is shown for simplicity. The disadvantages of the conventional approach are those of excessive size, weight, and power consumption, since the circuitry must be duplicated for each input carrier.

The disadvantages associated with processing each carrier separately can be avoided by means of *group processing*, in which the input FDMA signals are demultiplexed as a group in a single processing circuit, illustrated in Fig. 14.25*b*. Feasibility studies are being conducted into the use of digital-type group processors, although it would appear that these may require *very high speed integrated circuits* (VHSICs) not presently available. A different approach to the problem of group processing has been proposed, which makes use of an analog device known as a *surface acoustic wave* (SAW) Fourier transformer (Atzeni et al., 1975; Hays et al., 1975; Hays and Hartmann, 1976; Maines and Paige, 1976; Nud and Otto, 1975).

In its basic form, the SAW device consists of two electrodes deposited on the surface of a piezoelectric dielectric. An electrical signal applied to the input electrode sets up a SAW which induces a corresponding signal in the output electrode. In effect, the SAW device is a coupled circuit in which the coupling mechanism is the SAW.

Because the propagation velocity of the acoustic wave is much lower than that of an electromagnetic wave, the SAW device exhibits useful delay characteristics. In addition, the electrodes are readily shaped to

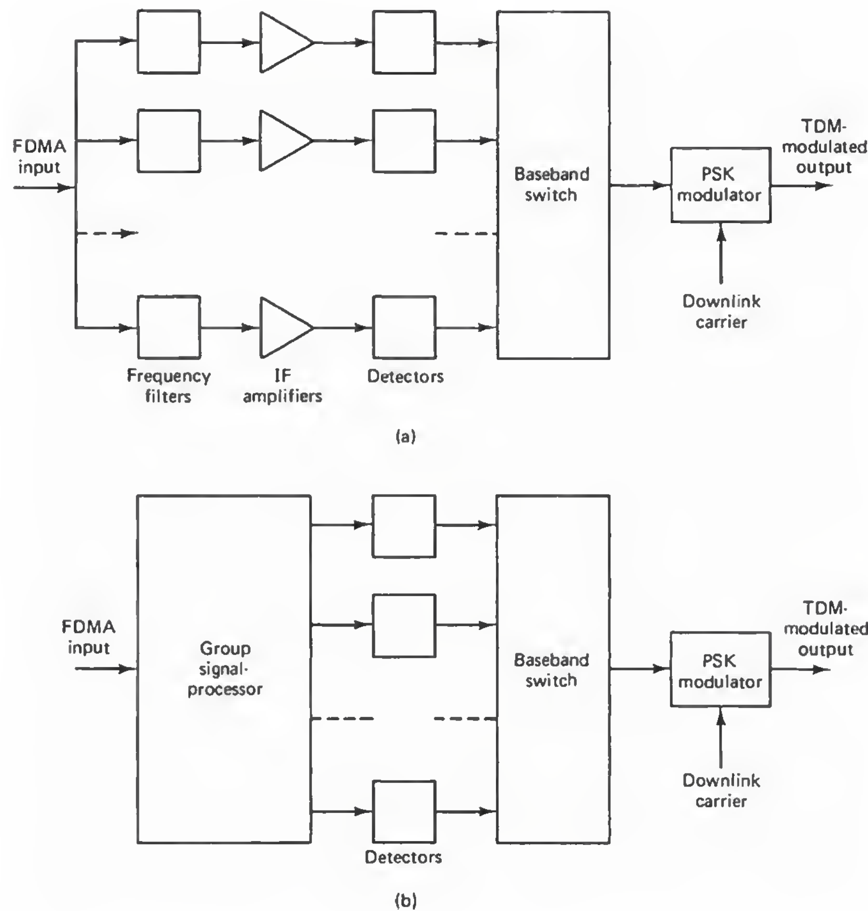


Figure 14.25 On-board signal processing for FDMA/TDM operation; (*a*) conventional approach; (*b*) group signal processing.

provide a wide range of useful transfer characteristics. These two features, along with the fact that the device is small, rugged, and passive, make it a powerful signal-processing component. SAW devices may be used conventionally as delay lines, as bandpass, or bandstop filters, and they are the key component in a unit known as a Fourier transformer.

The Fourier transformer, like any other transformer, works with input and output signals which are functions of time. The unique property of the Fourier transformer is that the output signal is a time analog of the frequency spectrum of the input signal. When the input is a group of FDMA carriers, the output in the ideal case would be an analog of the FDMA frequency spectrum. This allows the FDMA signals to be demultiplexed in real time by means of a commutator switch, which eliminates

the need for the separate frequency filters required in the conventional analog approach. Once the signals have been separated in this way, the original modulated-carrier waveforms may be recovered through the use of SAW inverse Fourier transformers.

In a practical transformer, continuous operation can only be achieved by repetitive cycling of the transformation process. As a result, the output is periodic, and the observation interval has to be chosen to correspond to the desired spectral interval. Also, the periodic interval over which transformation takes place results in a broadening of the output "pulses" which represent the FDMA spectra.

Repetitive operation of the Fourier transformer at a rate equal to the data bit rate will produce a suitable repetitive output. The relative positions of the output pulses will remain unchanged, fixed by the frequencies of the FDMA carriers. The PSK modulation on the individual FDMA carriers appears in the phase of the carriers within each output pulse. Thus the FDMA carriers have been converted to a pulsed TDM signal. Further signal processing is required before this can be retransmitted as a TDM signal.

Figure 14.26 shows the output obtained from a practical Fourier transformer for various input signals. For Fig. 14.26*a* the input was seven *continuous-wave* (CW) signals applied in succession. The output is seen to be pulses corresponding to the line spectra for these waves.

The broadening of the lines is a result of the finite time gate over which the Fourier transformer operates. It is important to note that the horizontal axis in Fig. 14.26 is a time axis on which the equivalent frequency points are indicated.

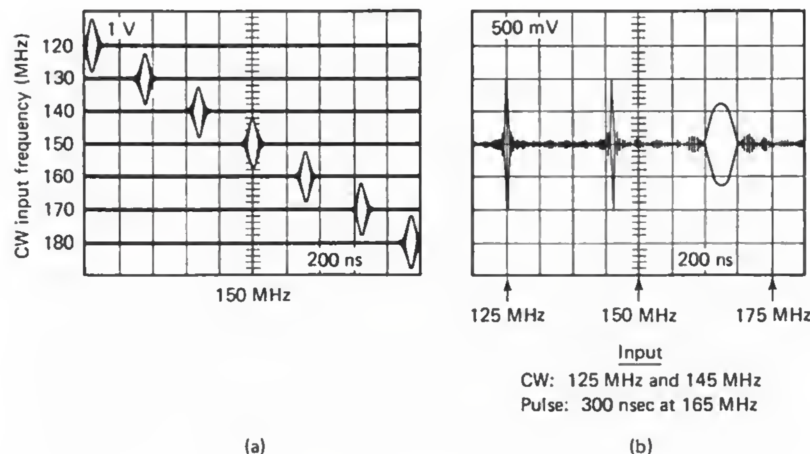


Figure 14.26 Prototype chirp transform of (*a*) seven successive CW input signals and (*b*) three simultaneous input signals, including CW and pulsed RF. 200 ns/div; 31.5 MHz/s chirp rate. (*Courtesy of Hays and Hartmann, 1976. Copyright 1976, IEEE.*)

Figure 14.26*b* shows the output obtained with three simultaneous inputs, two CW waves and one pulsed carrier wave. Again, the output contains two pulses corresponding to the CW signals and a time function which has the shape of the spectrum for the pulsed wave (Hays and Hartmann, 1976). A detailed account of SAW devices will be found in Morgan (1985) and in the IEEE Proceedings (1976).

14.9 Satellite-Switched TDMA

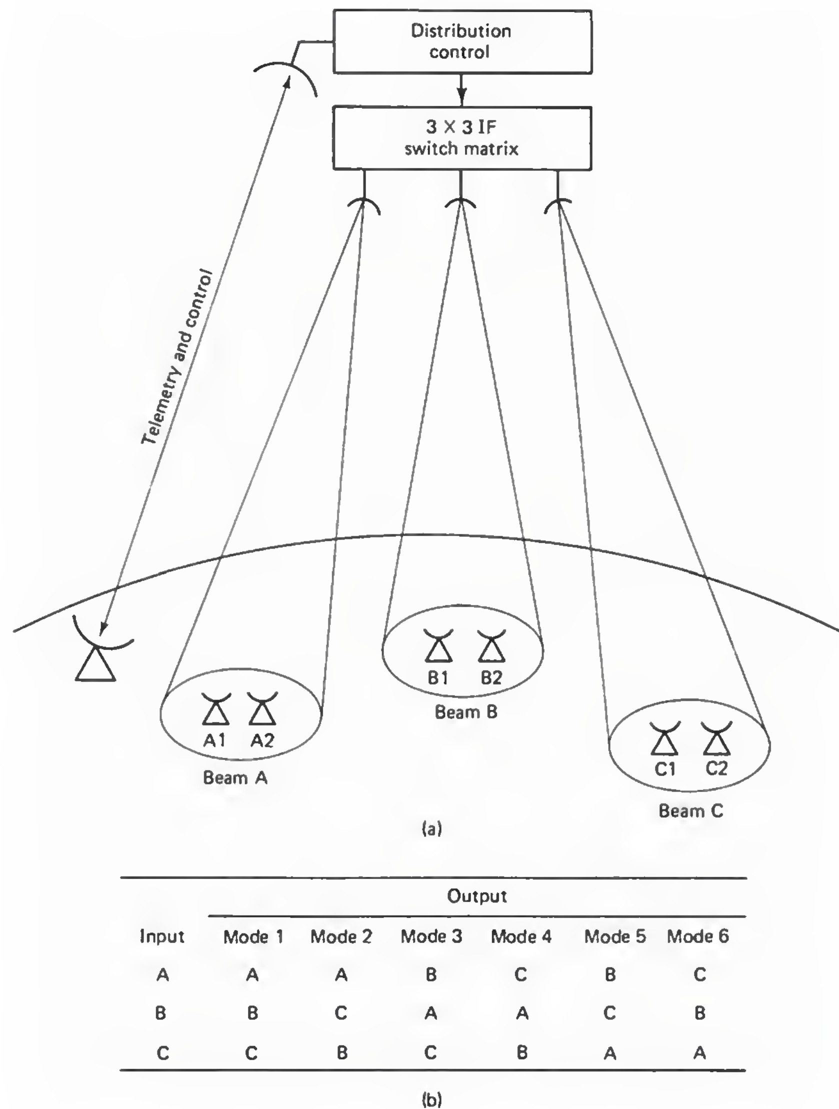
More efficient utilization of satellites in the geostationary orbit can be achieved through the use of antenna spot beams. The use of spot beams is also referred to as *space-division multiplexing.* Further improvements can be realized by switching the antenna interconnections in synchronism with the TDMA frame rate, this being known as *satellite-switched* TDMA (SS/TDMA).

Figure 14.27*a* shows in simplified form the SS/TDMA concept (Scarcella and Abbott, 1983). Three antenna beams are used, each beam serving two earth stations. A 3 × 3 satellite switch matrix is shown. This is the key component that permits the antenna interconnections to be made on a switched basis. A *switch mode* is a connectivity arrangement. With three beams, six modes would be required for full interconnectivity, as shown in Fig. 14.27*b*, and in general with N beams, $N!$ modes are required for full interconnectivity. Full interconnectivity means that the signals carried in each beam are transferred to each of the other beams at some time in the switching sequence. This includes the loopback connection, where signals are returned along the same beam, enabling intercommunications between stations within a beam. Of course, the uplink and downlink microwave frequencies are different.

Because of beam isolation, one frequency can be used for all uplinks, and a different frequency for all downlinks (e.g., 14 and 12 GHz in the Ku band). To simplify the satellite switch design, the switching is carried out at the intermediate frequency that is common to uplinks and downlinks. The basic block schematic for the 3 × 3 system is shown in Fig. 14.28.

A mode *pattern* is a repetitive sequence of satellite switch modes, also referred to as SS/TDMA frames. Successive SS/TDMA frames need not be identical, since there is some redundancy between modes. For example, in Fig. 14.27*b*, beam A interconnects with beam B in modes 3 and 5, and thus not all modes need be transmitted during each SS/TDMA frame. However, for full interconnectivity, the *mode pattern* must contain all modes.

All stations within a beam receive all the TDM frames transmitted in the downlink beam. Each frame is a normal TDMA frame consisting



	Output					
Input	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	Mode 6
A	A	A	B	C	B	C
B	B	C	A	A	C	B
C	C	B	C	B	A	A

(b)

Figure 14.27 (*a*) Satellite switching of three spot beams; (*b*) connectivities or modes.

of bursts, addressed to different stations in general. As mentioned, successive frames may originate from different transmitting stations and therefore have different burst formats. The receiving station in a beam recovers the bursts addressed to it in each frame.

The two basic types of switch matrix are the *crossbar matrix* and the *rearrangeable network*. The crossbar matrix is easily configured for the *broadcast mode*, in which one station transmits to all stations. The

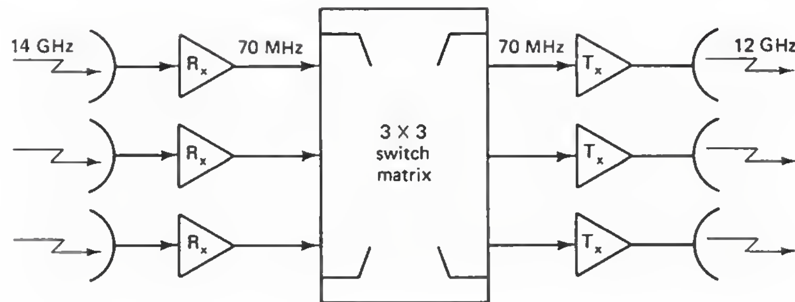


Figure 14.28 Switch matrix in the R.F. link.

broadcast mode with the rearrangeable network-type switch is more complex, and this can be a deciding factor in favor of the crossbar matrix (Watt, 1986). The schematic for a 3 × 3 crossbar matrix is shown in Fig. 14.29, which also shows input beam *B* connected in the broadcast mode.

The switching elements may be ferrites, diodes, or transistors. The dual-gate FET appears to offer significant advantages over the other types and is considered by some to be the most promising technology (Scarcella and Abbott, 1983).

Figure 14.30 shows how a 3 × 3 matrix switch may be used to reroute traffic. Each of the ground stations *U*, *V*, and *W* accesses a separate antenna on the satellite and carries traffic destined for the downlink

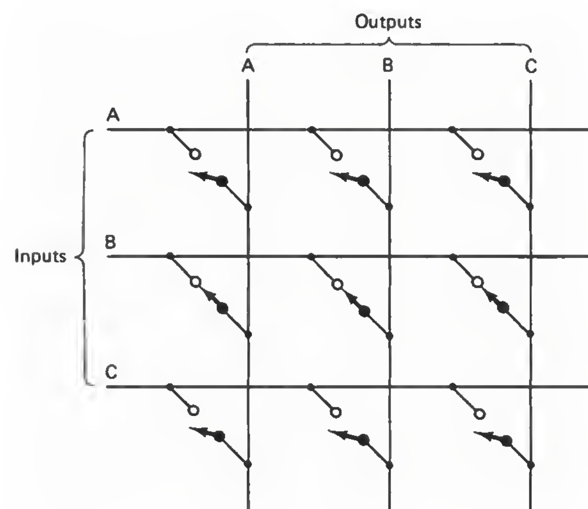


Figure 14.29 3 × 3 crossbar matrix switch, showing input *B* connected in the broadcast mode.

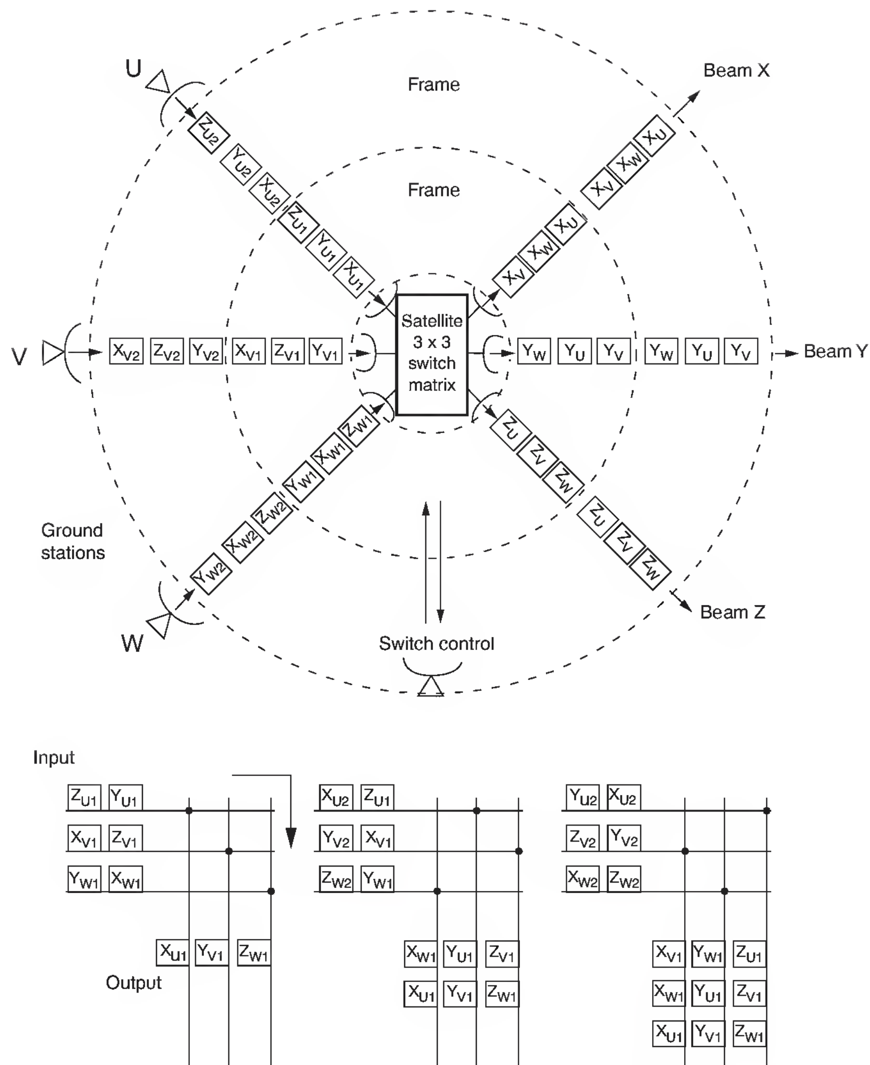


Figure 14.30 Traffic from earth stations U, V, W rerouted into designated beams X, Y, Z. The lower diagrams show part of the switching sequence.

beams *X*, *Y*, and *Z*. The switch is controlled from a ground control station, and the switching sequence for the frame labeled with subscript 1 is shown in the lower part of the figure.

The schematic for a 4 × 4 matrix switch as used on the European Olympus satellite is shown in Fig. 14.31 (Watt, 1986). This arrangement is derived from the crossbar matrix. It permits broadcast mode operation, but does not allow more than one input to be connected to one

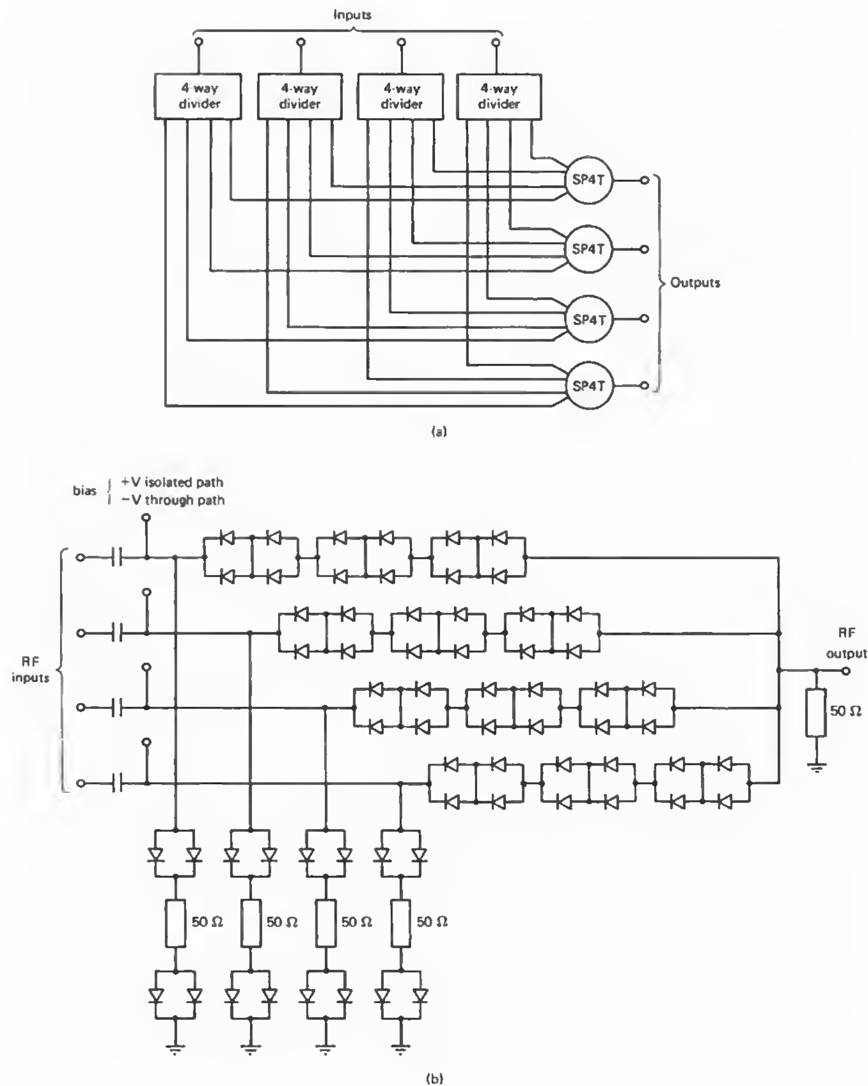


Figure 14.31 (*a*) 4 × 4 switch matrix; (*b*) circuit diagram of redundant SP4T switch element. (*Courtesy of Watt, 1986; reprinted with permission of IEE, London.*)

output. Diodes are used as switching elements, and as shown, diode quads are used which provide redundancy against diode failure. It is clear that satellite-switched TDMA adds to the complexity of the on-board equipment and to the synchronization requirements.

Use of multiple antenna beams can also be used for *space-division multiple access* (SDMA). Both beam switching, and the use of phased adaptive arrays (see Sec. 6.18) have been studied for mobile applications.

An analysis and comparison of antenna beam switching and phased adaptive arrays will be found in Zaharov (2001).

14.10 Code-Division Multiple Access

With CDMA the individual carriers may be present simultaneously within the same rf bandwidth, but each carrier carries a unique code waveform (in addition to the information signal) that allows it to be separated from all the others at the receiver. The carrier is modulated in the normal way by the information waveform and then is further modulated by the code waveform to spread the spectrum over the available rf bandwidth. Many of the key properties of CDMA rely on this spectrum spreading, and the systems employing CDMA are also known as *spread-spectrum multiple access* (SSMA). Care must be taken not to confuse the SS here with that for satellite switched (SS/TDMA) used in the Sec. 14.9.

CDMA can be used with analog and digital signals (see Dixon, 1984), but only digital systems will be described here. For illustration purposes, a polar *non-return-to-zero* (NRZ) waveform denoted by $p(t)$ (see Fig. 10.2) will be used for the information signal, and BPSK modulation (see Sec. 10.6.1) will be assumed. The code waveform $c(t)$ is also a polar NRZ signal, as sketched in Fig. 14.32. What would be called *bits* in an information waveform are called *chips* for the code waveform, and in most practical systems the chip rate is much greater than the information bit rate. The pulses (chips) in the code waveform vary randomly between $+V$ and $-V$. The randomness is an essential feature of spread-spectrum systems, and more will be said about this shortly. The code signal may be applied as modulation in exactly the same way as the information signal so that the BPSK signal carries both the information signal $p(t)$ and the code signal $c(t)$. This method is referred to as *direct-sequence spread spectrum* (DS/SS). Other techniques are also used to spread the spectrum, such as frequency hopping, but the discussion here will be limited to the DS/SS method.

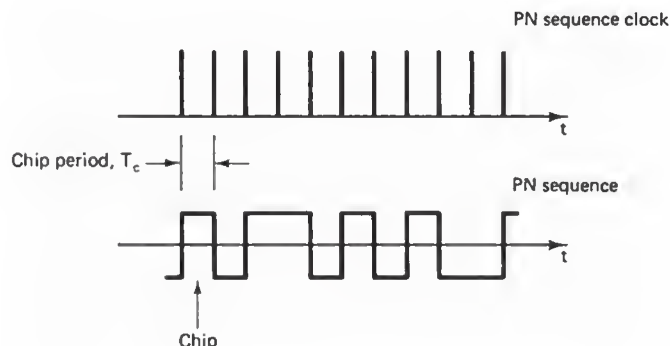


Figure 14.32 PN binary sequence. One element is known as a chip.

14.10.1 Direct-sequence spread spectrum

In Fig. 14.33, $p(t)$ is an NRZ binary information signal, and $c(t)$ is a NRZ binary code signal. These two signals form the inputs to a multiplier (balanced modulator), the output of which is proportional to the product $p(t)c(t)$. This product signal is applied to a second balanced modulator, the output of which is a BPSK signal at the carrier frequency. For clarity, it is assumed that the carrier is the uplink frequency, and hence the uplink carrier is described by

$$e_U(t) = c(t)p(t)\cos\omega_U t \tag{14.33}$$

The corresponding downlink carrier is

$$e_D(t) = c(t)p(t)\cos\omega_D t \tag{14.34}$$

At the receiver, an identical $c(t)$ generator is synchronized to the $c(t)$ of the downlink carrier. This synchronization is carried out in the *acquisition and tracking block*. With $c(t)$ a polar NRZ type waveform, and with the locally generated $c(t)$ exactly in synchronism with the transmitted $c(t)$, the product $c^2(t) = 1$. Thus the output from the multiplier is

$$\begin{aligned} c(t)e_D(t) &= c^2(t)p(t)\cos\omega_D t \\ &= p(t)\cos\omega_D t \end{aligned} \tag{14.35}$$

This is identical to the conventional BPSK signal given by Eq. (10.14), and hence detection proceeds in the normal manner.

14.10.2 The code signal *c*(*t*)

The code signal $c(t)$ carries a binary code that has special properties needed for successful implementation of CDMA. The binary symbols used in the codes are referred to as *chips* rather than *bits* to avoid confusion

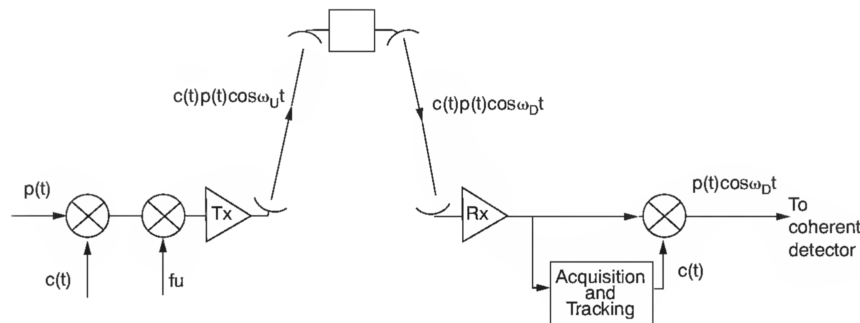


Figure 14.33 A basic CDMA system.

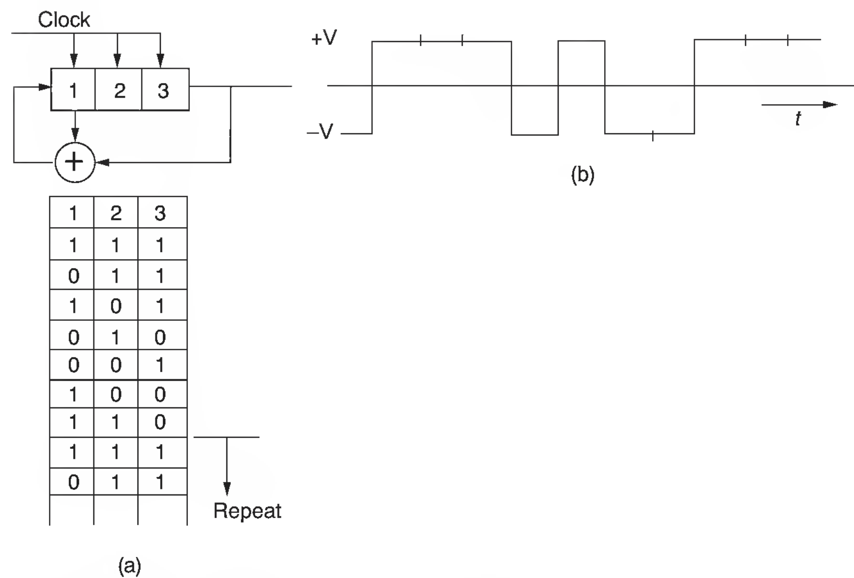
with the information bits that will also be present. Chip generation is controlled by a clock, and the chip rate, in chips per second, is given by the clock speed. Denoting the clock speed by R_{ch}, the chip period is the reciprocal of the clock speed:

$$T_{ch} = \frac{1}{R_{ch}} \tag{14.36}$$

The waveform $c(t)$ is periodic, in that each period is a repetition of a given sequence of N chips. The sequence itself exhibits random properties, which will be described shortly. The periodic time for the waveform is

$$T_N = NT_{ch} \tag{14.37}$$

The codes are generated using binary shift registers and associated linear logic circuits. The circuit for a three-stage shift register that generates a sequence of $N = 7$ chips is shown in Fig. 14.34*a*. Feedback occurs from stages 1 and 3 as inputs to the exclusive OR gate. This provides the input to the shift register, and the chips are clocked through at the clock rate R_{ch}. The generator starts with all stages holding binary 1s, and the following states are as shown in the table in Fig. 14.34. Stage 3 also provides the binary output sequence. The code waveform generated from this code is shown in Fig. 14.34*b*.



1	2	3
1	1	1
0	1	1
1	0	1
0	1	0
0	0	1
1	0	0
1	1	0
1	1	1
0	1	1

Figure 14.34 Generation of a 7-chip maximal sequence code.

Such codes are known as *maximal sequence* or *m-sequence* codes because they utilize the maximum length sequence that can be generated. For Fig. 14.34*a* the maximum length sequence is 7 chips as shown. In general, the shift register passes through all states (all combinations of 1s and 0s in the register) except the all-zero state when generating a maximal sequence code. Therefore, a code generator employing an n-stage shift register can generate a maximum sequence of N chips, where

$$N = 2^n - 1 \tag{14.38}$$

The binary 1s and 0s are randomly distributed such that the code exhibits noiselike properties. However, there are certain deterministic features described below, and the codes are more generally known as *PN codes*, which stands for *pseudo-noise codes*.

1. The number of binary 1s is given by

$$\text{No. of 1s} = \frac{2^n}{2} \tag{14.39}$$

and the number of binary 0s is given by

$$\text{No. of 0s} = \frac{2^n}{2} - 1 \tag{14.40}$$

The importance of this relationship is that when the code uses $+V$ volts for a binary 1 and $-V$ volts for a binary 0, the dc offset is close to zero. Since there is always one more positive chip than negative, the dc offset will be given by

$$\text{dc offset} = \frac{V}{N} \tag{14.41}$$

The dc offset determines the carrier level relative to the peak value; that is, the carrier is suppressed by amount $1/N$ for BPSK. For example, using a code with $n = 8$ with BPSK modulation, the carrier will be suppressed by 1/255 or 48 dB.

2. The total number of maximal sequences that can be generated by an n-stage shift register (and its associated logic circuits) is given by

$$S_{\max} = \frac{\phi(N)}{n} \tag{14.42}$$

Here, $\phi(N)$ is known as *Euler's ϕ-function*, which gives the number of integers in the range 1, 2, 3 . . ., $N - 1$, that are relatively prime to N

[N is given by Eq. (14.38)]. Two numbers are relatively prime when their greatest common divisor is 1. A general formula for finding $\phi(N)$ is (see Ore, 1988)

$$\phi(N) = N\left(\frac{p_1 - 1}{p_1}\right) \cdots \left(\frac{p_r - 1}{p_r}\right) \tag{14.43}$$

where $p_1 \ldots p_r$ are the prime factors of N. For example, for $n = 8$, $N = 2^8 - 1 = 255$. The prime factors of 255 are 3, 5, and 17, and hence

$$\phi(255) = 255\left(\frac{2}{3}\right)\left(\frac{4}{5}\right)\left(\frac{16}{17}\right)$$
$$= 128$$

The total number of maximal sequences that can be generated by an eight-stage code generator is therefore

$$S_{\text{max}} = \frac{128}{8} = 16$$

As a somewhat simpler example, consider the case when $n = 3$. In this instance, $N = 7$. There is only one prime factor, 7 itself, and therefore

$$\phi(7) = 7 \cdot \frac{6}{7} = 6$$

and

$$S_{\text{max}} = \frac{\phi(7)}{3} = 2$$

In this case there are only two distinct maximal sequences.

One of the most important properties of $c(t)$ is its *autocorrelation function*. The autocorrelation function is a measure of how well a time-shifted version of the waveform compares with the unshifted version. Figure 14.35*a* shows how the comparison may be made. The $c(t)$ waveform is multiplied with a shifted version of itself, $c(t - \tau)$, and the output is averaged (shown by the integrator). The average, of course, is independent of time t (the integrator integrates out the time-t dependence), but it will depend on the time lead or lag introduced by τ. When the waveforms are coincident, $\tau = 0$, and the average output is a maximum, which for convenience will be normalized to 1. Any shift in time, advance or delay, away from the $\tau = 0$ position will result in a decrease in output voltage. A property of m-sequence code waveforms is that the autocorrelation

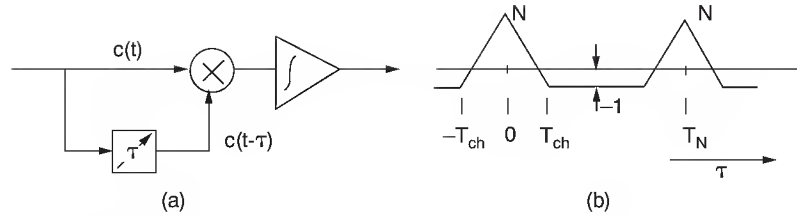


Figure 14.35 (*a*) Generating the autocorrelation function; (*b*) the autocorrelation waveform.

function decreases linearly from the maximum value (unity in this case) to a negative level $1/N$, as shown in Fig. 14.35*b*. The very pronounced peak in the autocorrelation function provides the chief means for acquiring acquisition and tracking so that the locally generated m-sequence code can be synchronized with the transmitted version.

14.10.3 Acquisition and tracking

One form of acquisition circuit that makes use of the autocorrelation function is shown in Fig. 14.36. The output from the first multiplier is

$$\begin{aligned} e(t) &= c(t - \tau)c(t)p(t)\cos\omega_D t \\ &= c(t - \tau)c(t)\cos[\omega_D t + \varphi(\mathrm{t})] \end{aligned} \tag{14.44}$$

Here, the information modulation, which is BPSK, is shown as $\varphi(t)$ so that the effect of the following *bandpass filter* (BPF) on the amplitude can be more clearly seen. The BPF has a passband centered on ω_D, wide with respect to the information modulation but narrow with respect to the code signal. It performs the amplitude-averaging function on the

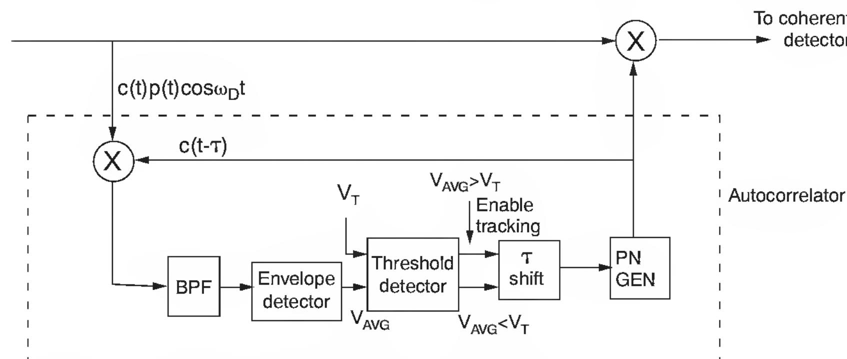


Figure 14.36 Acquisition of a carrier in a CDMA system.

code signal product (see Maral and Bousquet, 1998). The averaging process can be illustrated as follows. Consider the product of two cosine terms and its expansion:

$$\cos \omega t \cos(\omega t - \delta) = \frac{1}{2}\{\cos[\omega t + (\omega t - \delta)] + \cos[\omega t - (\omega t - \delta)]\}$$
$$= \frac{1}{2}[\cos(2\omega t - \delta) + \cos \delta] \tag{14.45}$$

The BPF will reject the high-frequency component, leaving only the average component 0.5 cos(δ). This signal may be considered analogous to the $c(t)c(t - \tau)$ term in Eq. (14.44). The envelope detector following the BPF produces an output proportional to the envelope of the signal, that is, to the average value of $c(t)c(t - \tau)$. This is a direct measure of the autocorrelation function. When it is less than the predetermined threshold V_T required for synchronism, the time shift τ incremented. Once the threshold has been reached or exceeded, the system switches from acquisition mode to tracking mode.

One form of tracking circuit, the *delay lock loop*, is shown in Fig. 14.37. Here, two correlators are used, but the local signal to one is advanced by half a chip period relative to the desired code waveform, and the other is delayed by the same amount. The outputs from the correlators are subtracted, and this difference signal provides the control voltage for the VCO that drives the shift register clock. With the control voltage at the zero crossover point, the locally generated code signal is in phase with the received code signal. Any tendency to drift out of phase changes the VCO in such a way as to bring the control voltage back to the zero crossover point, thus maintaining synchronism.

The acquisition and tracking circuits also will attempt to correlate the stored version of $c(t)$ at the receiver with all the other waveforms being received. Such correlations are termed *cross-correlations*. It is essential that the cross-correlation function does not show a similar peak as the autocorrelation, and this requires careful selection of the spreading functions used in the overall system (see, for example, Dixon, 1984).

14.10.4 Spectrum spreading and despreading

In Sec. 10.6.3 the idea of bandwidth for PSK modulation was introduced. In general, for a BPSK signal at a bit rate R_b, the main lobe of the power-density spectrum occupies a bandwidth extending from $f_c - R_b$ to $f_c + R_b$. This is sketched in Fig. 14.38*a*. A similar result applies when the

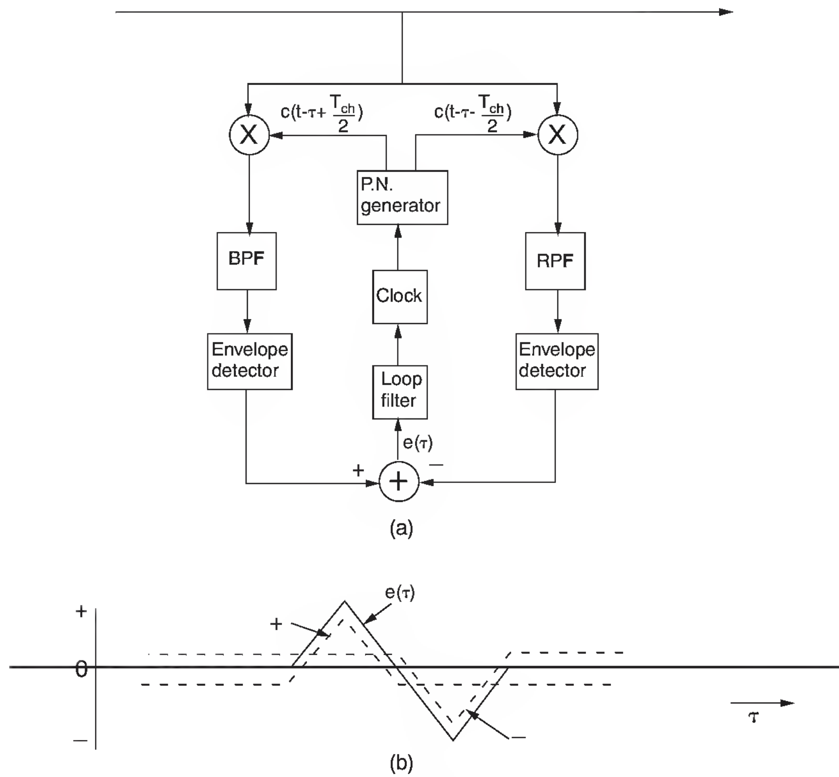


Figure 14.37 (*a*) The delay lock loop; (*b*) the waveform at the adder.

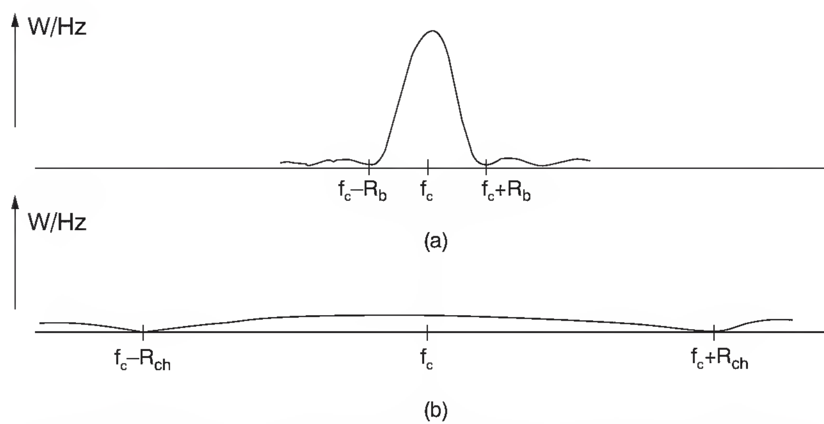


Figure 14.38 Spectrum for a BPSK signal: (*a*) without spreading, (*b*) with spreading.

modulation signal is $c(t)$, the power-density spectrum being as sketched in Fig. 14.38*b*. It should be mentioned here that because $c(t)$ exhibits periodicity, the spectrum density will be a line function, and Fig. 14.38*b* shows the envelope of the spectrum. The spectrum shows the *power density* (watts per hertz) in the signal. For constant carrier power, it follows that if a signal is forced to occupy a wider bandwidth, its spectrum density will be reduced. This is a key result in CDMA systems. In all direct-sequence spread-spectrum systems, the chip rate is very much greater than the information bit rate, or $R_{ch} >> R_b$. The bandwidth is determined mainly by R_{ch} so that the power density of the signal described by Eq. (14.34) is spread over the bandwidth determined by R_{ch}. The power density will be reduced approximately in the ratio of R_{ch} to R_b.

Assuming then that acquisition and tracking have been accomplished, $c(t)$ in the receiver (Fig. 14.33) performs in effect a *despreading function* that it restores the spectrum of the wanted signal to what it was before the spreading operation in the transmitter. This is also how the spread-spectrum technique can reduce interference. Figure 14.39*a* shows the spectra of two signals, an interfering signal that is not part of the CDMA system and that has not been spread, and the desired DS/SS received signal. Following the despreading operation for the desired signal, its spectrum is restored as described previously. The interfering signal, however, is simply multiplied by the $c(t)$ signal, which results in it being spread.

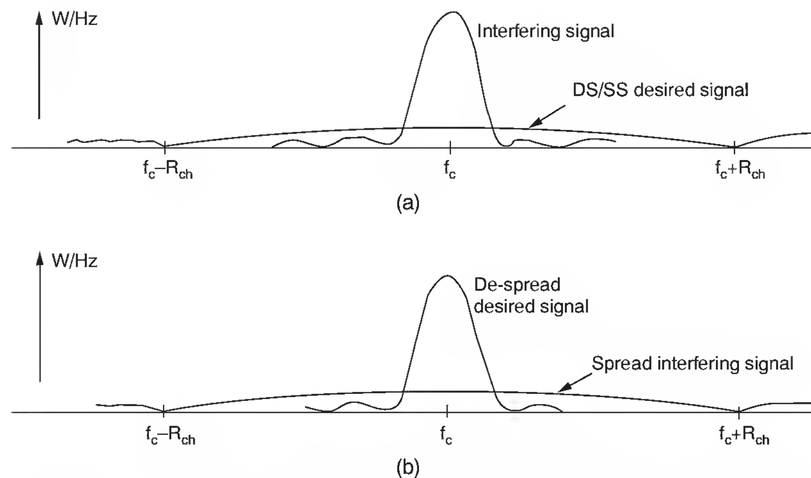


Figure 14.39 (*a*) Spectrum of an interfering, nonspread signal along with the spread desired signal; (*b*) the effect of the despreading operation on the desired signal resulting in spread of the interferer.

14.10.5 CDMA throughput

The maximum number of channels in a CDMA system can be estimated as follows: It is assumed that the thermal noise is negligible compared with the noise resulting from the overlapping channels, and also for comparison purposes, it will be assumed that each channel introduces equal power P_R into the receiver. For a total of K channels, $K - 1$ of these will produce noise, and assuming that this is evenly spread over the noise bandwidth B_N of the receiver, the noise density, in W/Hz, is

$$N_0 = \frac{(K - 1)P_R}{B_N} \tag{14.46}$$

Let the information rate of the wanted channel be R_b; then, from Eq. (10.22),

$$E_b = \frac{P_R}{R_b} \tag{14.47}$$

Hence the bit energy to noise density ratio is

$$\frac{E_b}{N_0} = \frac{B_N}{(K - 1)R_b} \tag{14.48}$$

The noise bandwidth at the BPSK detector will be approximately equal to the IF bandwidth as given by Eq. (10.15), but using the chip rate

$$\begin{aligned} B_N &\cong B_{\mathrm{IF}} \\ &= (1 + \rho)R_{\mathrm{ch}} \end{aligned} \tag{14.49}$$

where ρ is the rolloff factor of the filter. Hence the bit energy to noise density ratio becomes

$$\frac{E_b}{N_0} = \frac{(1 + \rho)R_{\mathrm{ch}}}{(K - 1)R_b} \tag{14.50}$$

As pointed out in Chap. 10, the probability of bit error is usually a specified objective, and this determines the E_b/N_0 ratio, for example, through Fig. 10.17. The number of channels is therefore

$$K = 1 + (1 + \rho)\frac{R_{\mathrm{ch}}}{R_b}\frac{N_0}{E_b} \tag{14.51}$$

The processing *gain* G_p is basically the ratio of power density in the unspread signal to that in the spread signal. Since the power density is inversely proportional to bandwidth, an approximate expression for the processing gain is

$$G_p = \frac{R_{\text{ch}}}{R_b} \tag{14.52}$$

Hence

$$K = 1 + (1 + \rho)G_p\frac{N_0}{E_b} \tag{14.53}$$

Example 14.8 The code waveform in a CDMA system spreads the carriers over the full 36 MHz bandwidth of the channel, and the rolloff factor for the filtering is 0.4. The information bit rate is 64 kb/s, and the system uses BPSK. Calculate the processing gain in decibels. Given that the BER must not exceed 10^{-5}, give an estimate of the maximum number of channels that can access the system.

Solution

$$\begin{aligned} R_{\text{ch}} &= \frac{B_{\text{IF}}}{1 + \rho} \\ &= \frac{36 \times 10^6}{1.4} \\ &= 25.7 \times 10^6 \text{ chips/s} \end{aligned}$$

Hence the processing gain is

$$\begin{aligned} G_p &= \frac{25.7 \times 10^6}{64 \times 10^3} \\ &= 401.56 \end{aligned}$$

From Fig. 10.18 for $P_e = 10^{-5}$, $[E_b/N_0] = 9.6$ dB approximately. This is a power ratio of 9.12, and from Eq. (14.53),

$$\begin{aligned} K &= 1 + \frac{1.4 \times 401.56}{9.12} \\ &\cong \underline{\underline{62 \text{ (rounded down)}}} \end{aligned}$$

The *throughput efficiency* is defined as the ratio of the total number of bits per unit time that can be transmitted with CDMA to the total number of bits per unit time that could be transmitted with single access and no spreading. For K accesses as determined earlier, each at bit rate R_b, the total bits per unit time is KR_b. A single access could utilize the full bandwidth, and hence its transmission rate as determined

by Eq. (10.15) is

$$R_T = \frac{B_{\text{IF}}}{1 + \rho} \tag{14.54}$$

This is the same as the chip rate, and hence the throughput is

$$\begin{aligned} \eta &= \frac{KR_b}{R_T} \\ &= \frac{KR_b}{R_{\text{ch}}} \\ &= \frac{K}{G_p} \end{aligned} \tag{14.55}$$

Using the values obtained in Example 14.8 gives a throughput of 0.15, or 15 percent. This should be compared with the frame efficiency for TDMA (see Example 14.4), where it is seen that the throughput efficiency can exceed 90 percent.

CDMA offers several advantages for satellite networking, especially where VSAT-type terminals are involved. These are:

1. The beamwidth for VSAT antennas is comparatively broad and therefore could be subject to interference from adjacent satellites. The interference rejection properties of CDMA through spreading are of considerable help here.
2. Multipath interference, for example, that resulting from reflections, can be avoided provided the time delay of the reflected signal is greater than a chip period and the receiver locks onto the direct wave.
3. Synchronization between stations in the system is not required (unlike TDMA, where synchronization is a critical feature of the system). This means that a station can access the system at any time.
4. Degradation of the system (reduction in E_b/N_0) is gradual with an increase in number of users. Thus additional traffic could be accommodated if some reduction in performance was acceptable.

The main disadvantage is the low throughput efficiency.

14.11 Problems and Exercises

14.1. Explain what is meant by a *single access* in relation to a satellite communications network. Give an example of the type of traffic route where single access would be used.

14.2. Distinguish between *preassigned* and *demand-assigned traffic* in relation to a satellite communications network.

14.3. Explain what is meant by FDMA, and show how this differs from FDM.

14.4. Explain what the abbreviation SCPC stands for. Explain in detail the operation of a preassigned SCPC network.

14.5. Explain what is meant by *thin route service*. What type of satellite access is most suited for this type of service?

14.6. Briefly describe the ways in which demand assignment may be carried out in an FDMA network.

14.7. Explain in detail the operation of the Spade system of demand assignment. What is the function of the common signaling channel?

14.8. Explain what is meant by *power-limited* and *bandwidth-limited operation* as applied to an FDMA network. In an FDMA scheme the carriers utilize equal powers and equal bandwidths, the bandwidth in each case being 5 MHz. The transponder bandwidth is 36 MHz. The saturation EIRP for the downlink is 34 dBW, and an output backoff of 6 dB is employed. The downlink losses are 201 dB, and the destination earth station has a G/T ratio of 35 dBK^{-1}. Determine the $[C/N]$ value assuming this is set by single carrier operation. Determine also the number of carriers which can access the system, and state, with reasons, whether the system is power limited or bandwidth limited.

14.9. A satellite transponder has a saturation EIRP of 25 dBW and a bandwidth of 27 MHz. The transponder resources are shared equally by a number of FDMA carriers, each of bandwidth 3 MHz, and each requiring a minimum EIRP of 12 dBW. If 7 dB output backoff is required, determine the number of carriers that can be accommodated.

14.10. In some situations it is convenient to work in terms of the carrier-to-noise temperature. Show that $[C/T] = [C/N_0] + [k]$. The downlink losses for a satellite circuit are 196 dB. The earth station $[G/T]$ ratio is 35 dB/K, and the received $[C/T]$ ratio is -138 dBW/K. Calculate the satellite [EIRP].

14.11. The earth-station receiver in a satellite downlink has an FM detector threshold level of 10 dB and operates with a 3-dB threshold margin. The emphasis improvement figure is 4 dB, and the noise-weighting improvement figure is 2.5 dB. The required $[S/N]$ at the receiver output is 46 dB. Calculate the receiver processing gain. Explain how the processing gain determines the IF bandwidth.

14.12. A 252-channel FM/FDM telephony carrier is transmitted on the downlink specified in Prob. 14.10. The peak/rms ratio factor is 10 dB, and the baseband bandwidth extends from 12 to 1052 kHz. The voice-channel bandwidth

is 3.1 kHz. Calculate the peak deviation, and hence, using Carson's rule, calculate the IF bandwidth.

14.13. Given that the IF bandwidth for a 252-channel FM/FDM telephony carrier is 7.52 MHz and that the required [*C*/*N*] ratio at the earth-station receiver is 13 dB, calculate (*a*) the [*C*/*T*] ratio and (*b*) the satellite [EIRP] required if the total losses amount to 200 dB and the earth-station [*G*/*T*] ratio is 37.5 dB/K.

14.14. Determine how many carriers can access an 80-MHz transponder in the FDMA mode, given that each carrier requires a bandwidth of 6 MHz, allowing for 6.5-dB output backoff. Compare this number with the number of carriers possible without backoff.

14.15. (*a*) Analog television transmissions may be classified as *full-transponder* or *half-transponder transmissions*. State what this means in terms of transponder access. (*b*) A composite TV signal (video plus audio) has a top baseband frequency of 6.8 MHz. Determine for a 36-MHz transponder the peak frequency deviation limit set by (1) half-transponder and (2) full transponder transmission.

14.16. Describe the general operating principles of a TDMA *network*. Show how the transmission bit rate is related to the input bit rate.

14.17. Explain the need for a reference burst in a TDMA system.

14.18. Explain the function of the preamble in a TDMA traffic burst. Describe and compare the channels carried in a preamble with those carried in a reference burst.

14.19. What is the function of (*a*) the burst-code word and (*b*) the carrier and bit-timing recovery channel in a TDMA burst?

14.20. Explain what is meant by (*a*) *initial acquisition* and (*b*) *burst synchronization* in a TDMA network. (*c*) The nominal range to a geostationary satellite is 42,000 km. Using the station-keeping tolerances stated in Sec. 7.4 in connection with Fig. 7.10, determine the variation expected in the propagation delay.

14.21. (*a*) Define and explain what is meant by *frame efficiency* in relation to TDMA operation. (*b*) In a TDMA network the reference burst and the preamble each requires 560 bits, and the nominal guard interval between bursts is equivalent to 120 bits. Given that there are eight traffic bursts and one reference burst per frame and the total frame length is equivalent to 40,800 bits, calculate the frame efficiency.

14.22. Given that the frame period is 2 ms and the voice-channel bit rate is 64 kb/s, calculate the equivalent number of voice channels that can be carried by the TDMA network specified in Prob. 14.21.

14.23. Calculate the frame efficiency for the CSC shown in Fig. 14.19.

14.24. (*a*) Explain why the frame period in a TDMA system is normally chosen to be an integer multiple of 125 μs. (*b*) Referring to Fig. 14.20 for the INTELSAT preassigned frame format, show that there is no break in the timing interval for sample 18 when this is transferred to a burst.

14.25. Show that, all other factors being equal, the ratio of uplink power to bit rate is the same for FDMA and TDMA. In a TDMA system the preamble consists of the following slots, assigned in terms of number of bits: bit timing recovery 304; unique word 48; station identification channel 8; order wire 64. The guard slot is 120 bits, the frame reference burst is identical to the preamble, and the burst traffic is 8192 bits. Given that the frame accommodates 8 traffic bursts, calculate the frame efficiency. The traffic is preassigned PCM voice channels for which the bit rate is 64 kb/s, and the satellite transmission rate is nominally 60 Mb/s. Calculate the number of voice channels which can be carried.

14.26. In comparing design proposals for multiple access, the two following possibilities were considered: (1) uplink FDMA with downlink TDM and (2) uplink TDMA with downlink TDM. The incoming baseband signal is at 1.544 Mb/s in each case and the following table shows values in decilogs:

	Uplink	Downlink
$[E_b/N_0]$	12	12
$[G/T]$	10	19.5
[LOSSES]	212	210
[EIRP]	—	48
Transmit Antenna Gain $[G_T]$	45.8	—

Determine (*a*) the downlink TDM bit rate and (*b*) the transmit power required at the uplink earth station for each proposal.

14.27. A TDMA network utilizes QPSK modulation and has the following symbol allocations: guard slot 32; carrier and bit timing recovery 180; burst code word (unique word) 24; station identification channel 8; order wire 32; management channel (reference bursts only) 12; service channel (traffic bursts only) 8. The total number of traffic symbols per frame is 115,010, and a frame consists of two reference bursts and 14 traffic bursts. The frame period is 2 ms. The input consists of PCM channels each with a bit rate of 64 kb/s. Calculate the frame efficiency and the number of voice channels that can be accommodated.

14.28. For the network specified in Prob. 14.27 the BER must be at most 10^{-5}. Given that the receiving earth-station $[G/T]$ value is 30 decilogs and total losses are 200 dB, calculate the satellite [EIRP] required.

14.29. Discuss briefly how demand assignment may be implemented in a TDMA network. What is the advantage of TDMA over FDMA in this respect?

14.30. Define and explain what is meant by the terms *telephone load activity factor* and *digital speech interpolation*. How is advantage taken of the load activity factor in implementing digital speech interpolation?

14.31. Define and explain the terms *connect clip* and *freeze-out* used in connection with digital speech interpolation.

14.32. Describe the principles of operation of a SPEC system, and state how this compares with digital speech interpolation.

14.33. Determine the bit rate that can be transmitted through a 36-MHz transponder, assuming a rolloff factor of 0.2 and QPSK modulation.

14.34. On a satellite downlink, the $[C/N_0]$ ratio is 86 dBHz and an $[E_b/N_0]$ of 12 dB is required at the earth station. Calculate the maximum bit rate that can be transmitted.

14.35. FDMA is used for uplink access in a satellite digital network, with each earth station transmitting at the T1 bit rate of 1.544 Mb/s. Calculate (*a*) the uplink $[C/N_0]$ ratio required to provide a $[E_b/N_0] = 14$ dB ratio at the satellite and (*b*) the earth station [EIRP] needed to realize the $[C/N_0]$ value. The satellite $[G/T]$ value is 8 dB/K, and total uplink losses amount to 210 dB.

14.36. In the satellite network of Prob. 14.35, the downlink bit rate is limited to a maximum of 74.1 dBb/s, with the satellite TWT operating at saturation. A 5-dB output backoff is required to reduce intermodulation products to an acceptable level. Calculate the number of earth stations that can access the satellite on the uplink.

14.37. The [EIRP] of each earth station in an FDMA network is 47 dBW, and the input data are at the T1 bit rate with 7/8 FEC added. The downlink bit rate is limited to a maximum of 60 Mb/s with 6-dB output backoff applied. Compare the [EIRP] needed for the earth stations in a TDMA network utilizing the same transponder.

14.38. (*a*) Describe the general features of an on-board signal processing transponder that would allow a network to operate with FDMA uplinks and a TDMA downlink. (*b*) In such a network, the overall BER must not exceed 10^{-5}. Calculate the maximum permissible BER of each link, assuming that each link contributes equally to the overall value.

14.39. Explain what is meant by *full interconnectivity* in connection with satellite switched TDMA. With four beams, how many switch modes would be required for full interconnectivity?

14.40. Identify all the redundant modes in Fig. 14.27.

14.41. The shift register in an *m*-sequence generator has 7 stages. Calculate the number of binary 1s and 0s. The code is used to generate a NRZ polar

waveform at levels +1 V and −1 V. Calculate the dc offset and the carrier suppression in decibels that can be achieved when BPSK is used.

14.42. The shift register in an m-sequence generator has 10 stages. Calculate the length of the m-sequences. Determine the prime factors for N and, hence, the total number of maximal length sequences that can be produced.

14.43. As shown in Sec. 14.10.2, an m-sequence generator having a 3-stage shift register is capable of generating a total of 2 maximal sequences, and Fig. 14.34 shows one of these. Draw the corresponding circuit for the other sequence, and the waveform.

14.44. Draw accurately to scale the autocorrelation function over one complete cycle for the waveform shown in Fig. 14.34. Assume $V = 1$ V and $T_{ch} = 1$ ms.

14.45. Draw accurately to scale the autocorrelation function over one complete cycle for the waveform determined in Prob. 14.43. Assume $V = 1$ V, and $T_{ch} = 1$ ms.

14.46. Describe in your own words how signal acquisition and tracking are achieved in a DS/SS system.

14.47. An m-sequence generator having a 3-stage shift register is capable of generating a total of 2 maximal sequences. Neatly sketch the cross-correlation function for the two m-sequences.

14.48. Explain the principle behind spectrum spreading and despreading and how this is used to minimize interference in a CDMA system.

14.49. The IF bandwidth for a CDMA system is 3 MHz, the rolloff factor for the filter being 1. The information bit rate is 2.4 kb/s, and an $[E_b/N_0]$ of 11 dB is required for each channel accessing the CDMA system. Calculate the maximum number of accesses permitted.

14.50. Determine the throughput efficiency for the system in Prob. 14.49.

14.51. Show that when K is large such that the first term, unity, on the right hand side of Eq. (14.53) can be neglected, the throughput efficiency is independent of the processing gain. Hence, plot the throughput efficiency as a function of $[E_b/N_0]$ for the range 7 to 11 dB.

References

Atzeni, C., G. Manes, and L. Masotti. 1975. "Programmable Signal Processing by Analog Chirp-Transformation using SAW Devices." *Ultrasonics Symposium Proceedings*, IEEE, New York.

CCIR Report 708 (mod I). 1982. "Multiple Access and Modulation Techniques in the Fixed Satellite Service." *14th Plenary Assembly*, Geneva.

Dixon, R. C. 1984. *Spread Spectrum Systems.* Wiley, New York.

Freeman, R. L. 1981. *Telecommunications Systems Engineering*. Wiley, New York.
Gagliardi, R. M. 1991. *Satellite Communications*, 2d ed. Van Nostrand Reinhold, New York.
Ha, T. T. 1990. *Digital Satellite Communications*, 2d ed. McGraw-Hill, New York.
Hays, R. M., W. R. Shreve, D. T. Bell, Jr., L. T. Claiborne, and C. S. Hartmann. 1975. "Surface Wave Transform Adaptable Processor System." *Ultrasonics Symposium Proceedings*, IEEE, New York.
Hays, Ronald M., and C. S. Hartmann. 1976. "Surface Acoustic Wave Devices for Communications." *Proc. IEEE*, Vol. 64, No. 5, May, pp. 652–671.
IEEE Proceedings. 1976. Special Issue on Surface Acoustic Waves. May
INTELSAT. 1980. "Interfacing with Digital Terrestrial Facilities." ESS-TDMA-1-8, p. 11.
Lewis, J. R. 1982. "Satellite Switched TDMA. Colloquium on the Global INTELSAT VI Satellite System." *Digest No. 1982/76*, IEEE, London.
Maines, J. D., and E. G. S. Paige. 1976. "Surface Acoustic Wave Devices for Signal Processing Applications." *Proc. IEEE*, Vol. 64, No. 5, May.
Maral, G., and M. Bousquet. 1998. *Satellite Communications Systems*. Wiley, New York.
Martin, J. 1978. *Communications Satellite Systems*. Prentice-Hall, Englewood Cliffs, NJ.
Miya, K. (ed.). 1981. *Satellite Communications Technology*. KDD Engineering and Consulting, Inc., Japan.
Morgan, D. P. 1985. *Surface-Wave Devices for Signal Processing*. Elsevier, New York.
Nudd, G. R., and O. W. Otto. 1975. "Chirp Signal Processing Using Acoustic Surface Wave Filters." *Ultrasonics Symposium Proceedings*, IEEE, New York.
Nuspl, P. P., and R. de Buda. 1974. "TDMA Synchronization Algorithms." *IEEE EASCON Conference Record*, Washington, DC.
Ore, O. 1988. *Number Theory and Its History*. Dover Publications, New York.
Pratt, T., and C. W. Bostian. 1986. *Satellite Communications*. Wiley, New York.
Rosner, R. D. 1982. *Packet Switching*. Lifetime Learning Publications, New York.
Scarcella, T., and R. V. Abbott. 1983. "Orbital Efficiency Through Satellite Digital Switching." *IEEE Communications Magazine*, Vol. 21, No. 3, May.
Sciulli, J. A., and S. J. Campanella. 1973. "A Speech Predictive Encoding Communication System for Multichannel Telephony." *IEEE Transactions on Communications*, Vol. Com-21, No. 7, July.
Spilker, J. J. 1977. *Digital Communications by Satellite*. Prentice-Hall, Englewood Cliffs, NJ.
Stevenson, S., W. Poley, L. Lekan, and J. Salzman. 1984. "Demand for Satellite-Provided Domestic Communications Services to the Year 2000." NASA Tech. Memo. 86894, November.
Watt, N. 1986. "Multibeam SS-TDMA Design Considerations Related to the Olympus Specialised Services Payload." *IEE Proc.*, Vol. 133, Part F, No. 4, July.
Zaharov, V. F. C., M. Gutierrez, 2001. "Smart Antenna Application for Satellite Communications Systems with Space Division Multiple Access." *Journal of Radio Electronics* N2 2001, at http://jre.cplire.ru/jre/feb01/1/text.html#fig1

Chapter

15

Satellites in Networks

15.1 Introduction

The word *network* has a variety of meanings, for example, in electrical circuits a network is an interconnection of two or more circuit elements such as resistors, inductors, capacitors, amplifiers and oscillators. If the network contains at least one closed path it is called a circuit (Hayt et al., 1978). In telecommunications work, a network is a connection of devices such as telephones, computers, switches, and printers. (And of course people can "network" together, which gives yet another meaning to the word).

Of particular interest to this chapter are *broadband networks*. The word *broadband* in the context of telecommunication networks means that the network is designed to carry voice, data, video, and image type signals. The *Internet* is a broadband network. This mixture of signals can also be carried by a method known as *asynchronous transfer mode* (ATM). In this chapter the word *network* will mean a broadband network.

A key feature of broadband networks is the way in which information is assembled into *packets* which can be transmitted on an "as needed" basis. This differs from previous methods where information was (and still is in many cases) transmitted as a continuous signal. Packet transmission as originally designed for data (such as computer output and email) was found to be unsuited for voice transmission, the packets being too large. With mixed packets of voice and data, the large data packets could introduce unacceptable delays in the reassembly of the voice packets. The ATM, which uses relatively small packets, known as *cells*, was designed specifically for broadband traffic (voice, data, video, images). The ATM is described in Sec. 15.4.

15.2 Bandwidth

Bandwidth is a key concept in any telecommunications network. Originally, the word bandwidth was used to define a range of frequencies and it applied mainly to analog circuits and signals, see, for example Sec. 9.6.2. Digital signals also require frequency bandwidth, as shown in Sec. 10.5, and here the frequency bandwidth is directly proportional to the bit rate. For example, as shown by Eq. (10.13), under certain conditions the bandwidth of a T1 signal is equal numerically to the bit rate. With broadband signals it is common practice to state the bandwidth as a bit rate.

The term *wideband* is also encountered in telecommunications networks, for example the bandwidth at the input to a satellite receiver is wideband, as described in Sec. 7.7.1. Although "wideband" and "broadband" would appear to be similar in meaning, each are used in quite specific contexts, wideband referring to the frequency range of signals and systems, and broadband as a bit rate as encountered in networks.

15.3 Network Basics

The *topology* of a network refers to the way in which the network devices are connected. For example, a *bus* is where the connections are strung out in a line. A *star* is where a number of connections radiate out from a hub. A *ring* is where the connection forms a ring, starting and ending at the same point, and the network devices are connected as spurs to the ring. A *node* is where a number of connections meet at a common point within the network. Two types of nodes are encountered, *terminal nodes* and *communicating nodes* (see, e.g., Walrand, 1991). A terminal node, as the name suggests is where terminal equipment such as computers and telephones, connect into the network. Communicating nodes are nodes within the network where various switching and routing functions are carried out. Nodes in general are complex structures, and in order to ensure the smooth flow of information through the node, the equipment at the node must operate to well defined rules, or what is referred to as a *protocol*. The overall protocol for a node is structured in *layers* where each layer has its own protocol (i.e., a set of rules) that is independent of the protocol in other layers.

Figure 15.1 shows how users A and B might communicate through a three-layer protocol. Each layer interfaces with the layers immediately adjacent, above and below. It is assumed that the information to be sent from A to B is in binary form. This may be a bit stream as generated, for example, by pulse code modulation for voice (see Sec. 10.3), or it could be bits stored in a digital file such as an email message. Originally "digital networks" in the context used here were separate from the "telephone network" and the information to be transferred was called *data*,

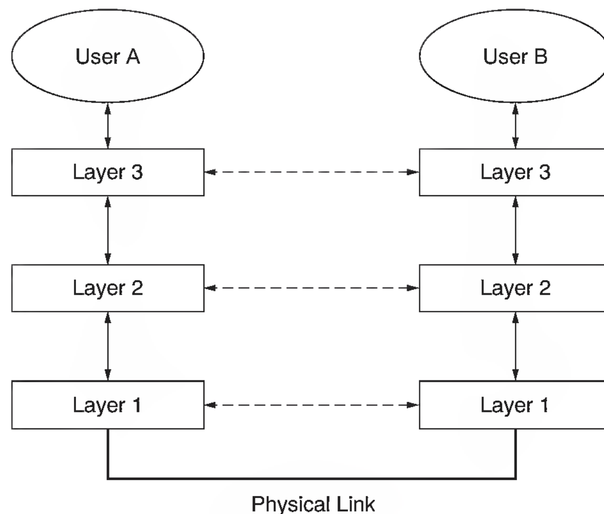


Figure 15.1 Network layers.

since it was mostly computer generated. However, as mentioned in the introduction, present day networks carry voice, data, video, and image type signals, all in digital form. Layer 3 interfaces with user A and may, for example, rearrange the bits into *packets*. There has to be *interface control information* exchanged between user A and layer 3, but once the message information has been transferred to layer 3, the packet arrangement in layer 3 does not depend on the user A. Layer 3 also interfaces with layer 2. Again, some interface control information is required to ensure a smooth transfer of the message information, now in packets in this example. Layer 2 might, for example, add more bits to the message to provide an address, or perhaps some form of error control, but this can be done without reference to layer 3. Layer 2 interfaces with layer 1, which is always known as the *physical layer* in network literature. The protocol in the physical layer takes care of such matters as specifying the type of connectors to be used, the type of modulation to be employed, and the nature of the physical link. In the case of satellites the physical link will be the equipment and properties of the uplink and downlink.

In Figure 15.1 dotted connections are shown between peer layers, each layer on user A's side is connected to the corresponding layer on user B's side. This is a *virtual connection* not a physical connection. The only physical connection is that through layer 1. Peer layers have to exchange information to ensure that the layer protocol is being followed, but such information is only of use to the two peer layers.

Some of the networks in common use are:

Local Area Network (LAN). Connects devices that are close geographically, for example in the same building.

Wide Area Network (WAN). Connects devices that are well-separated geographically, for example, long distance telephone lines or radio may have to be used for the connections.

Metropolitan Area Network (MAN). Designed for use in a town or city.

Campus Area Network (CAN). Designed for operation in a campus such as a military campus or the campus of an educational establishment.

Home Area Network (HAN). Designed to connect together devices in a person's home (e.g., computers and printers).

In this chapter the use of satellite communications in the *Internet*, probably the best known of networks, and in networks using the ATM will be examined.

Information may be transferred through a network in one of two ways, *connection oriented* and *connectionless oriented*. In connection oriented transfer, the path between sender and receiver is established before the information is transferred, and the information packets are transmitted in sequence. In connectionless oriented transfer a path may not be established before sending the packets, and these may not be sent in sequence. The packets in this case may follow different paths through the network. As an analogy, the connection oriented transfer could be compared to the exchange of information that takes place during a normal telephone conversation, while the connectionless oriented transfer is rather like sending information through the (snail) mail services.

15.4 Asynchronous Transfer Mode (ATM)

When information is transferred through a network it passes through a number of stages. The information is assembled into *packets*, and in broadband networks the packets are multiplexed to form a single bit stream. Then there is the actual transmission of signals from node to node, and at the nodes the signals will undergo some form of switching. The complete process of getting information from source to destination is referred to as a *transfer mode*. With the ATM, the packets originating from an individual user do not have to be transmitted at periodic intervals. This is what is meant by *asynchronous*. For comparison, the time division multiplexing described in Sec. 10.4 is a form of *synchronous* transmission. Asynchronous transfer mode is commonly denoted by ATM, and the "packets" are known as *cells*. Cells can be given time slots on demand as and when required.

15.4.1 ATM layers

The protocol for ATM was created by a number of standards organizations:

The International Telecommunication Union—Telecommunications Standardization Sector (ITU-T)

The American National Standards Institute (ANSI)

European Telecommunications Standards Institute (ETSI)

The ATM Forum

The layers for ATM are shown in Figure 15.2.

- *Physical layer.* The lowest layer is the physical layer (corresponding to layer 1 in Figure 15.1). The physical layer has two sublayers, the *physical medium sublayer*, which deals with such matters as line codes to be used and bit timing. Although the cells are transmitted asynchronously, bit timing is required for correct reception of the bit stream. This is provided at the physical medium sublayer. The other, (upper) sublayer is the *transmission convergence sublayer*. Certain transmission systems require the bit stream to be framed, (an example of framing is the T1 system shown in Figure 10.7). Packing cells into a frame, and unpacking them on receive, is carried out in the transmission convergence sublayer. The transmission conversion sublayer also provides the mechanism for delineating cells, that is identifying the boundaries of the cell. Also, for certain types of bursty data

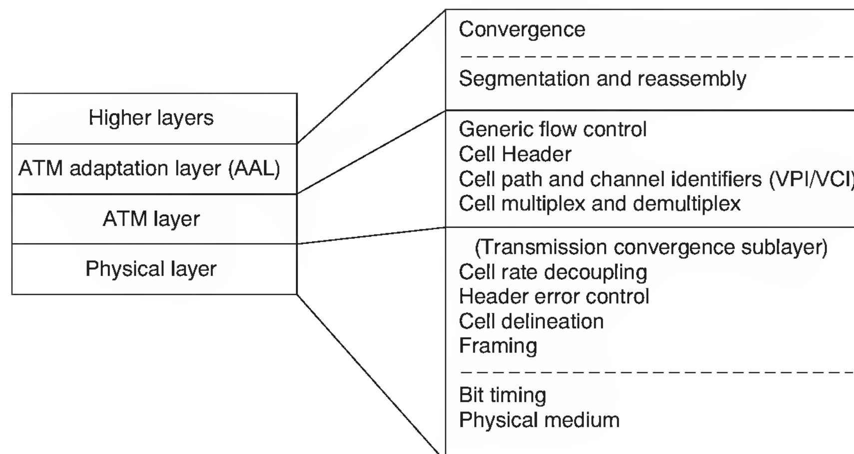


Figure 15.2 ATM layers.

such as computer output where there may be relatively long idle periods. "Idle" cells are inserted to maintain timing on transmit, and removed on reception at the transmission convergence sublayer (*cell decoupling*). Cell header error control is implemented at this transmission sublayer. The cell header is described in detail shortly.

- *ATM layer.* Cells from various sources (voice, data, video, images) are multiplexed, that is, arranged in sequence for transmit, and are separated, (de-multiplexed) into their respective channels on receive. Also at this layer a cell header is added which provides path and channel identification. The function of the header is key to ATM operation and this will be discussed in detail shortly.
- *ATM Adaptation Layer (AAL).* The adaptation layer is divided into two sub-layers. The *convergence sublayer* (CS) where the service requirements for the multimedia (voice, data, video, image) are established, and the *segmentation and reassembly sublayer* where the incoming information is segmented into cells, and the outgoing information is reassembled into its original format. The incoming bit streams for the various media will be quite different in bit rate and burstiness. The ATM adaptation layers are:
  - *AAl-1* is used for *constant bit rate* (CBR) applications, and is designed for voice and data that has to be sent over circuit facilities such as T1 (described in Sec. 10.4). As shown below 48 octets in a cell are reserved for the payload, but AAL-1 uses one of these octets for overhead, leaving 47 octets for the information payload.
  - *AAl-2* is used for *variable bit rate* (VBR) such as video and voice that use compression techniques. Again, part of the payload, from 1 to 3 octets is used for certain overhead functions, leaving 45 to 47 octets for the information payload.
  - *AAL-3/4*—Originally there was a separate AAL-3 but it was found that this could be combined with AAL-4. In this way AAL-3/4 was established. It is connectionless oriented (see Sec. 15.3) and is used for VBR data. An overhead of 4 octets is included in the payload field leaving 44 octets for the information payload. It should be noted that all the other AALs are connection oriented (see Sec. 15.3).
  - *AAL-5* is used for VBR and is connection oriented. The payload does not contain any overhead, so the information payload makes use of the available 48 octets.

There is also an ATM adaptation layer known as the null layer, and designated *AAL-0*. This is used where the information is already assembled in cell format.

15.4.2 ATM networks and interfaces

Two basic types of ATM networks are in general use—*public* and *private*. The public ATM network is designed to provide connections between any two subscribers, rather in the way that the public telephone network does. A private ATM network is one set up by an organization to provide ATM communications between the various parts of the organization. A private ATM network can exist independently of the public network, and the methods of interfacing to these will also differ in general.

A number of different interfaces are encountered in an ATM network, and these are denoted as follows:

- *User Network Interface* (UNI). This is the interface between an end user and an ATM network. It occurs at the entry point for a user. A UNI may be public or a private.
- *Network Node Interface* (NNI). This is the interface between two nodes in a network, which may be public of private. A private NNI is usually indicated by PNNI in the literature.
- *Network Network Interface*. This is also abbreviated NNI and is an interface between two ATM networks. The interface between two private networks is indicated by PNNI. The context should make clear in both instances whether the NNI refers to network-node or network-network interface. A private network has to interface with the public network through a public UNI.

15.4.3 The ATM cell and header

The standard ATM cell is made up of 53 octets, where an octet is a sequence of 8 bits. The 8-bit sequence is also known as a *byte*, but it should be noted that in computer terminology a byte can consist of other than 8 bits (see also Sec. 15.7). The cell structure shown in Fig. 15.3*a* is seen to consist of a 5 octet header, and a 48 octet payload. The header contains a number of *fields* which provide the information necessary to guide the cell through the network. These fields are described later, but as seen in Fig. 15.3, the header for cells at the point of entry to a network (at a UNI) differs slightly from the header for cells traversing a NNI. The header fields are:

GFC. This is the *generic flow control* field. Its function is to provide control and metering of the data flow before it enters the network. No flow control is exercised once the cell has entered the network. The 4 bits in the header are reallocated as described under VPI.

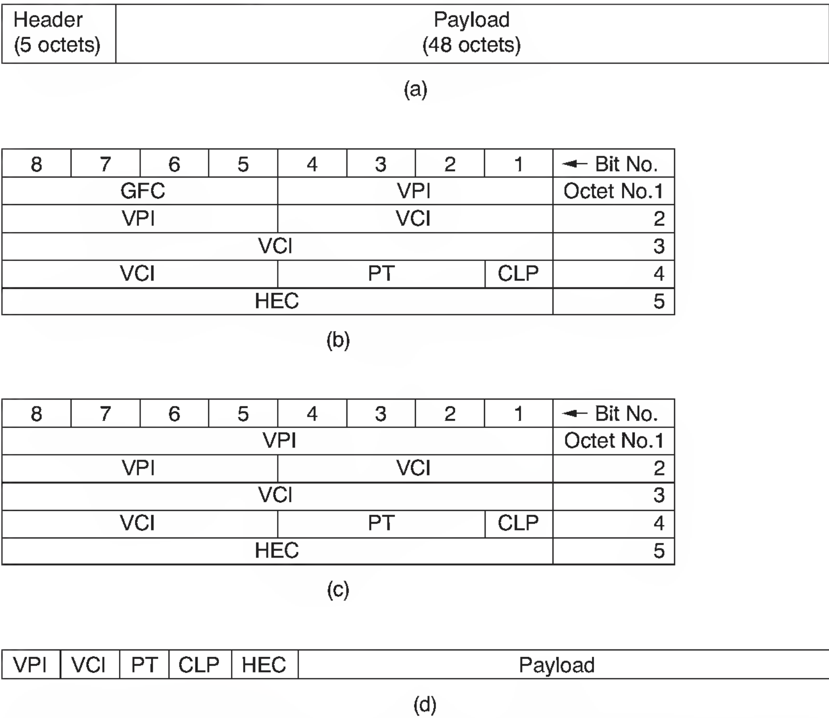


Figure 15.3 (*a*) ATM cell structure; (*b*) UNI header; (*c*) NNI header; (*d*) header and payload bit stream.

VCI. This is the *virtual channel identifier* field. A virtual channel carries a single stream of cells, which may be a mixture of voice, video, image, and data. From the point of view of the users, the different signals appear to have their own channels, and hence these are called *virtual channels*. Each channel is identified by its VCI.

VPI. This is the *virtual path identifier* field. Virtual channels can be "bundled" together to form what is termed a *virtual path*. With binary coding it should be kept in mind that all the information forms a single serial bit stream, which is transmitted over a physical link (e.g., optical fiber and satellite). The VPI enables the cells to be grouped into separate (but virtual) paths over the physical link. Switching within the network can be carried out by switching the virtual paths, without the need to switch the channels separately. As shown in Fig. 15.3, the GFC field is not required for cells once they are in the network, and the 4 bits for this field are reassigned to the VPI field, which enables the number of virtual paths within the network to be increased.

CLP. This is the *cell loss priority* field. It consists of a single bit, a 1 for a low priority cell, meaning that the cell can be discarded in the event of congestion. A 0 indicates high priority, meaning that the cell should only be discarded if it cannot be delivered.

HEC. This is the *header error control* field. The HEC field enables single bit errors in the header (including the HEC field) to be corrected, and double bit errors to be detected (but not corrected). The receiver is normally in the error correction mode, and if a single error is detected it will be corrected and the cell transmitted onward. However, after detection and correction of a single error the receiver automatically switches to the error detection mode, and of course if more than one error is detected to start with, it will automatically switch to error detection mode and the cell will be discarded. In error detection mode, *no errors* are corrected, and cells with errors are discarded. Once a cell with an error-free header is detected, the receiver automatically switches back to error-correction mode. The rationale behind this approach is discussed at length in Goralski (1995). Error control is applied only to the header (the data in the payload may have its own error-control coding). ATM was originally used for transmission over low *bit error rate* (BER) links such as optical fiber, where the bit error probability can be as low as 10^{-11} compared to a BER that can be as high as 10^{-2} over satellite links. Furthermore, bit errors in satellite links often occur in bursts which require special error-control coding (see Secs. 11.3.3 and 11.6). The HEC field is placed at the end of the header as shown in Fig. 15.3*d* where it functions also as a marker for cell position.

15.4.4 ATM switching

Keep in mind that the octets in the header form a serial bit stream when being transmitted, as shown in Fig. 15.3*d*. A cell is switched through a network on a path determined by the VPI and VCI fields. At each node in the network a cell is switched from one physical link to another. Two basic types of ATM switches are encountered. In one type, known as an ATM *digital cross connect switch* (DCS) the VPI field in a cell is overwritten by a new number on switching, while the VCI is left unchanged. This enables a number of virtual channels to be switched together, thus speeding up the switching process. The functioning of the DCS is illustrated in Fig. 15.4*a*, where for simplicity only the VPIs and VCIs are shown. These have been given hypothetical numbers for illustration purposes, for example 200/12 indicates a VCI of 200 and a VPI of 12. The look-up table in the DCS determines the new VPI for the virtual paths, and also the physical output link to which the paths must

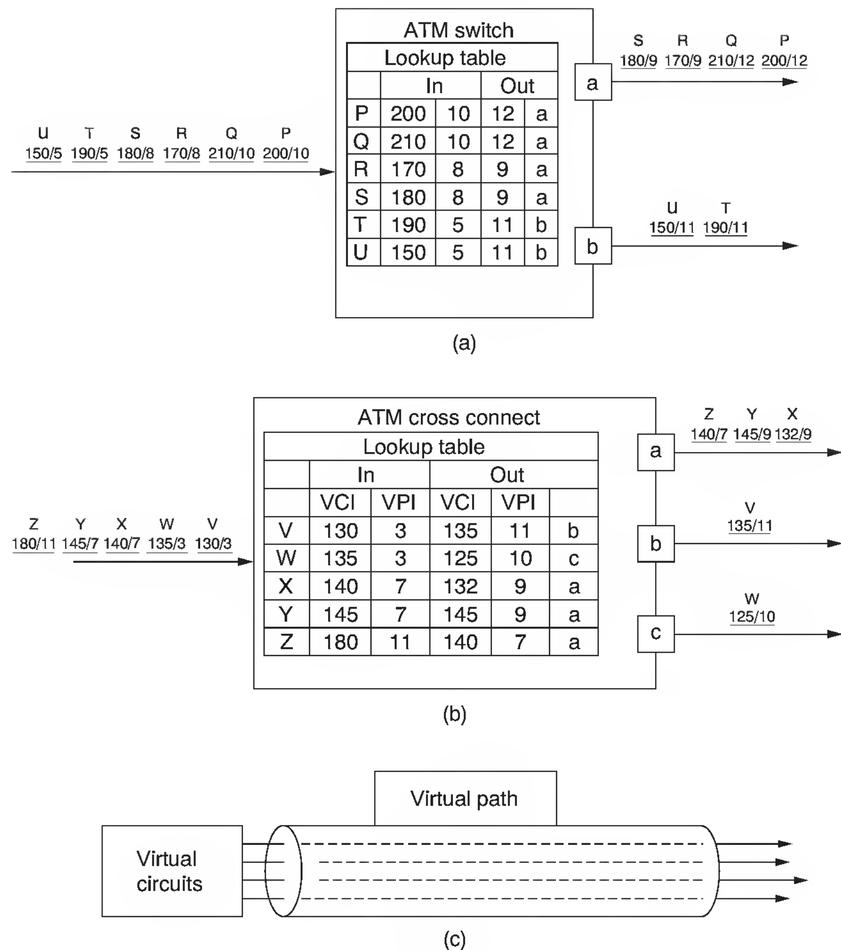


Figure 15.4 (*a*) ATM switch; (*b*) ATM cross connect; (*c*) virtual circuits and path.

be switched. Different suppliers offer different types of switching stages (the switch fabric). For ease of identification the cells have been labeled P, Q, R, S, T, U.

The second basic type of switch is referred to as an ATM switch. This differs from the DCS in that both the VPI and VCI fields are overwritten with new numbers on switching. Figure 15.4*b* illustrates the functioning of an ATM switch, where the VPI and the VCI of cells are both overwritten by new numbers based on the look-up table. Again, for ease of identification the cells have been labeled V, W, X, Y, Z. Although the virtual channels and virtual paths are formed over a single physical link, it is common practice to show the virtual channels bundled into a virtual path as in Fig. 15.4*c*.

From Fig. 15.3 it is seen that the VPI field at the UNI has 8 bits, and therefore 2^8 or 256 virtual paths can be supported. At the NNI the VPI field is increased to 12 bits, allowing 2^{12} or 4096 virtual paths to be supported. The VCI has 16 bits and therefore in theory the number of virtual channels is 2^{16} or 65536. This in practice is limited to 64000 (see Russell, 2000).

15.4.5 Permanent and switched virtual circuits

A virtual path may be set up on a permanent basis. The switching identifiers are preset so that the path does not change, and is available for use as required. This is referred to as a *permanent virtual circuit*, abbreviated PVC (note the use of the word "circuit" to describe this). A path may also be established anew each time a connection is required. The users are disconnected when the transfer is finished, and a new connection will be required if another session is asked for. This is described as a *switched virtual circuit* (SVC)— the term *switch* being used in analogy with telephone usage. The terms PVC and SVC can also be applied to virtual channels.

15.4.6 ATM bandwidth

A major advantage claimed for ATM is that it can provide bandwidth on demand, also known as *flexible bandwidth allocation.* The concept of bandwidth and its relationship to bit rate is described in Sec. 15.2. With the transmission of digital signals over networks the significant parameter is the bit rate. It is common practice, therefore, to state bandwidth as a bit rate. The maximum bit rate that a physical link can support is referred to as the *speed* of the link. Denoting the speed by S, the *maximum* bandwidth available for the payload can be calculated simply as S multiplied by the ratio of payload bits per cell to total bits per cell. The payload has 48 octets and hence the payload bits per cell is $8 \times 48 = 384$. An ATM cell has a total of 53 octets and therefore there is a total of $8 \times 53 = 424$ bits per cell. (The situation is somewhat more complicated as in certain cases the 48 octets contain some overhead bits, but this will be ignored for the present calculation). If a single user makes use of all the cells and these are transmitted without a break at speed S, then the payload bandwidth would be:

$$\mathrm{BW} = S \times \frac{384}{424} \tag{15.1}$$

For example, at a speed of 30 Mbps the payload bandwidth is approximately 27 Mbps.

Although it is possible for all the cells to be assigned to a single channel, it is much more likely that the cells will be distributed over many channels, that is, a number of signals will be *multiplexed* into the ATM transmission. This is one of the strengths of ATM, the ability to accommodate multiple signals, each requiring different bandwidths in general. The bandwidth used by a signal can be calculated from knowledge of the cell distribution. In practice cells may be transmitted in *frames*. Although the cells are asynchronous (that is, not necessarily distributed periodically over the frames), the frames themselves are transmitted synchronously, and in what is termed the *synchronous transmission module-1* (STM-1), used for ATM, 8000 frames are transmitted per second (see Mackenzie 1998). If one cell per frame is assigned to one user, and since each cell carries 384 payload bits, the corresponding bandwidth is:

$$B_{\text{payload}} = 384 \times F \tag{15.2}$$

where F is the number of frames per second. For example using the value $F = 8000$ frames/s, the payload bandwidth for one cell per frame is:

$$\begin{aligned} B_{\text{payload}} &= 384 \times 8000 \\ &= 3.072 \text{ Mbps} \end{aligned}$$

The flexibility of ATM arises because a cell may be sent only once every *nth* frame, or more than one cell per frame may be sent. Suppose for example a user cell is sent once every 48th frame, and the frame rate is 8000 frames/s then the corresponding payload bandwidth would be:

$$\begin{aligned} \frac{3.072}{48} &= 0.064 \text{ Mbps} \\ &= 64 \text{ kbps} \end{aligned}$$

As shown by Eq. (10.6) this is the bit rate for a single PCM signal.

If more than one cell per frame is allocated to a user, say C cells per frame then the corresponding payload bandwidth is

$$\begin{aligned} B_C &= B_{\text{payload}} \times C \\ &= 384 \times F \times C \end{aligned}$$

There is of course a limit to the number of cells that can fit in a frame. If S is the speed of the link in bits/s, as before, and F the number of frames/s, then the bits per frame is S/F. Each cell has 424 bits (the header

bits must be included here), and hence maximum number of cells per frame is

$$C_{\max} = \frac{S}{424 \times F} \tag{15.3}$$

Substituting this for C in the B_C equation and simplifying gives:

$$\mathrm{BW}_{C\max} = \frac{384}{424} S \tag{15.4}$$

This of course is just the Eq. 15.1 arrived at earlier for the maximum payload bandwidth for a single user making use of all the cells.

Voice, television, and videoconferencing all require a CBR service (essentially a fixed bandwidth). These signals are sensitive to variations in cell delay, so the cells for these services have to be spaced at regular intervals. Other services such as email and file transfer are not sensitive to variations in cell delay, and cells can be interspersed as space permits. This is called VBR service. The use of ATM for CBR and VBR is illustrated in Fig. 15.5.

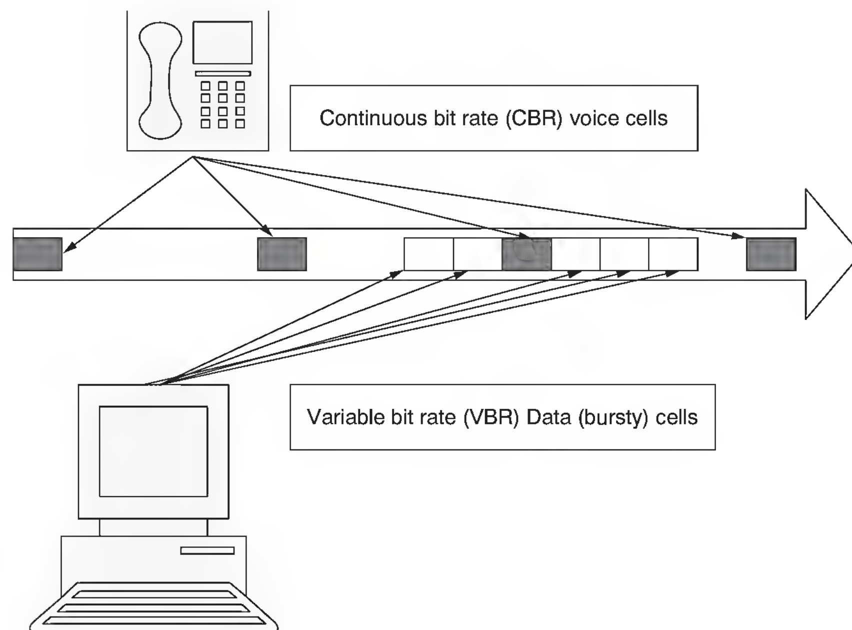


Figure. 15.5 The use of ATM for CBR and VBR.

TABLE 15.1 ATM Service Classes

Service class	Quality of service parameter
Constant bit rate (CBR)	This class is used for emulating circuit switching. The cell rate is constant with time, and applications are sensitive to cell-delay variation. Telephone traffic, videoconferencing, television are all examples of CBR transmissions.
Variable bit rate–non real time (VBR–NRT)	Traffic can be sent at a variable rate, which depends on the availability of user information. Statistical multiplexing* is used to optimize network resources. Multimedia is an example of VBR-NRT transmission.
Variable bit rate–real time (VBR–RT)	Similar to VBR–NRT, but designed for applications that are sensitive to cell delay. Application examples are voice with speech activity detection and interactive compressed video.
Available bit rate (ABR)	Provides rate-based flow control, the rate depending on the cell congestion present in the network. Examples of use are file transfer and email.
Unspecified bit rate (UBR)	A "catch all" class, widely used for TCP/IP (see Sec. 15.6)

NOTES: With statistical multiplexing the sum of the peak input rates exceeds the capacity of the link, but the assumption is made that sources are statistically independent and are unlikely to send peak rates at the same time.
SOURCE: With minor modifications from Web ProForum Tutorials, ©The International Engineering Consortium, http://www.iec.org.

15.4.7 Quality of service

In addition to flexible bandwidth allocation, ATM offers a guaranteed *quality of service* (QoS) which differs for various applications. For example voice and video are sensitive to cell delay and more so to variations in cell delay, but may be able to tolerate some loss of cells. Data on the other hand is not affected by cell delay (within limits) but cannot tolerate cell loss. When a connection is requested by a user the network is first tested to ensure that the guaranteed QoS can be provided, otherwise the user request is not accepted. There are in fact five service classes based on bit rate and these are summarized in Table 15.1.

A number of parameters are used in the measurement of ATM performance, and these are shown in Table 15.2.

15.5 ATM over Satellite

The ATM, described in the previous sections is used in terrestrial networks, to carry a mixture of signals, for example voice, data, video, and images. A natural development is to extend the ATM networks to include satellite links to bring the services of ATM to remote or isolated users, and also to provide broadcast facilities. Satellite links have the additional advantage of enabling *local area networks* (LANs) that are widely

TABLE 15.2 ATM Technical Parameters

Technical parameter	Definition
Cell loss ratio (CLR)	Percentage of cells not delivered to their destination because of congestion or buffer overflow.
Cell transfer delay (CTD)	The delay experienced by a cell between entry and exit points of a network
Cell delay variation (CDV)	A measure of the variance of the cell transfer delay.
Peak cell rate (PCR)	The maximum cell rate at which a user will transmit.
Sustained cell rate (SCR)	The average transmission rate measured over a long period, of the order of the connection lifetime.
Burst tolerance (BT)	Determines the maximum burst that can be sent at the peak rate.

SOURCE: With minor modifications from Web ProForum Tutorials, ©The International Engineering Consortium, http://www.iec.org.

separated geographically to be linked together, thus forming a *wide area network* (WAN). ATM over satellite is usually abbreviated as SATM in the literature (for satellite ATM). Satellites may be incorporated into ATM networks in a number of ways, some of which are described here, but there are certain problems unique to satellite links that have to be addressed in all cases.

The first of these is the BER. ATM was originally designed to operate over optical fiber links where bit errors are randomly distributed and the probability of bit error is comparatively low, in the order of 10^{-10}. In satellite links the BER is generally much higher than this, (see Figs. 10.17 and 11.8), and more importantly, bit errors can occur in bursts. In laboratory tests involving MPEG-2 signals it has been shown (Ivancic et al., 1997; 1998) that a BER of better than 10^{-8} is required for the most stringent ATM applications. Such values may be achieved using concatenated convolution and *Reed-Solomon* (R-S) *forward error correction* (FEC), (see Sec. 11.6).

The problem with applying FEC to ATM signals overall is that error correction is being applied to channels that might be tolerant of a higher BER such as voice, and thus is a waste of resources. A number of commercial error-control schemes for ATM over satellite are available. The Ericsson CLA-2000 ATM Link Accelerator (Ericsson, 2003) uses adaptive R-S and interleaving mechanisms, resulting in a *cell loss ratio* (CLR) and a *cell error ratio* (CER) of 10^{-10} or better (see Table 15.2). By adaptive is meant that the coding scheme is adjusted to suit the application (e.g., voice or data), and it dynamically adapts to the link conditions, for example, on clear days it uses less FEC overhead, achieving a 7 percent increase in bandwidth.

Another approach is taken in the LANET protocol, where LANET stands for *Limitless ATM Network*, a product of Yurie-Lucent. Adaptive R-S coding is applied to the payload and the PTI and CLP fields in the header (Akyildiz et al., 1998). Yurie have also implemented a redundancy coding scheme for the circuit (channel) addresses in the header. Multiple addresses are provided and are chosen such that the most probable error occurrences will change a given address to another permissible address within the group for the channel.

Delay, and variance of delay (termed *delay jitter*) also present problems not normally found in terrestrial networks. For geostationary satellites the one-way propagation delay is about 250 milliseconds. Variations in delay, termed *delay jitter*, can be more of a problem for delay sensitive channels such as voice and video, and some form of buffering is needed to minimize the delay jitter. The use of *low earth orbit* (LEO) and *medium earth orbit* (MEO) satellites reduces the propagation delay, although other problems are then introduced. LEO and MEO networks are discussed shortly.

ATM satellite networks can be broadly classified as *bent pipe architecture* and *on-board processing architecture*. With the "bent pipe" architecture, the satellite acts as a conduit between two earth stations, which may be fixed or mobile. With the *on-board processing* (usually abbreviated to OBP) architecture, ATM switches form part of the transponder in the satellite, which allows for on-board switching based on the VPI and VCI fields. This may be incorporated with the OBP and satellite beam switching described in Sec. 14.8 and 14.9. The use of OBP provides greater connectivity, reduces transmission delay, and permits the use of smaller and cheaper user terminals, but of course all this is at the cost of a more complex satellite structure.

The simplest situation is the bent pipe *relay architecture* illustrated in Fig. 15.6. This makes use of geostationary satellites, where the satellite link can be thought of as replacing a terrestrial link between two fixed points. The satellite link is likely to operate at a lower bit rate than the terrestrial ATM services connecting through it and some rate adaptation will be necessary. This is provided by the modems shown in Fig. 15.6. The *ATM link accelerator* (ALA), shown in Fig. 15.6, provides the adaptive error-control coding mentioned earlier. Cell switching (using the VPI and VCI fields) is carried out at the ATM end stations, not in the satellite. Also, signaling (e.g., call set up and teardown) takes place between the ground ATM switches. The relay designation applies only to "bent pipe" satellites, and is not applicable to OBP satellites.

The broad classifications, "bent pipe" and OBP defined earlier can be further classified as *network access*, and *network interconnect*. The distinction

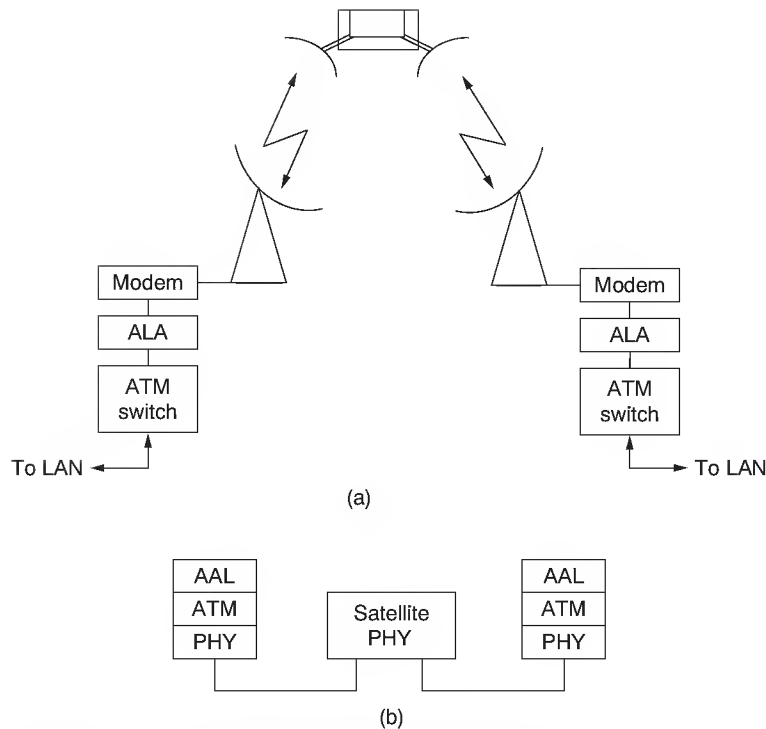


Figure 15.6 (*a*) "Bent pipe" satellite relay; (*b*) layer architecture.

here is that network access provides satellite access for ATM end-users, who may connect to each other and to ATM networks; network interconnect provides for satellite interconnection of ATM networks. Access and interconnect can be used with fixed and mobile users and networks (a mobile network for example could be a group of users on a ship or plane), where mobile operations must be supported. Fig. 15.7 shows the bent pipe access arrangement for fixed end-users and an ATM network. In general, the access and interconnect modes of operation may utilize a combination of *geostationary earth orbiting satellites* (GEOs), LEO, and MEO satellites.

The most complex arrangement, which applies only to OBP satellites, is called *full mesh*. Here the satellites themselves form an ATM network in space, in which the full complement of ATM functions operate. This includes traffic switching, flow and congestion control, connection setup and teardown, and quality of service requirements.

Where mobile operations have to be supported, mobility management functions are provided by a *mobility enhanced UNI layer* at the end

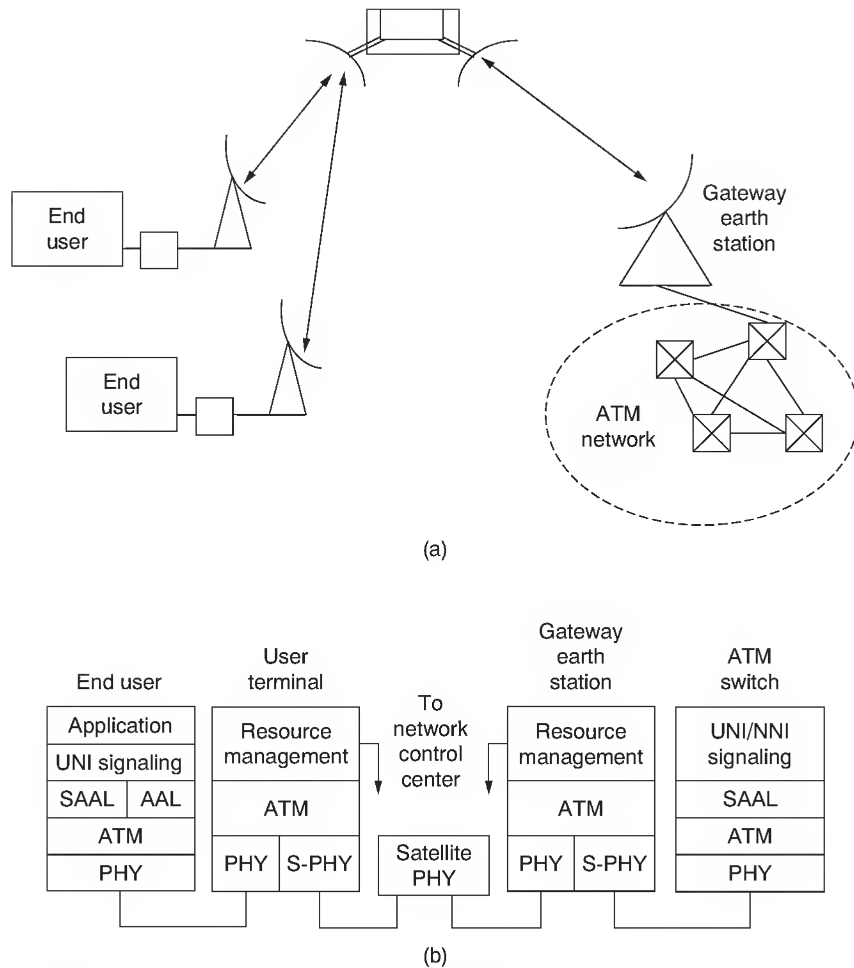


Figure 15.7 (*a*) Fixed network access; (*b*) layer architecture.

user (M+UNI) and by a mobility enhanced UNI/NNI layer (M+UNI/NNI) at the ATM switch. Mobility management includes *location management*, (authentication, registration, paging, roaming, and routing) and *handoff*, required when the mobile units move from one location to another.

An extensive overview of satellite architectures will be found in Toh and Li, (1998), a summary of which is given in Table 15.3.

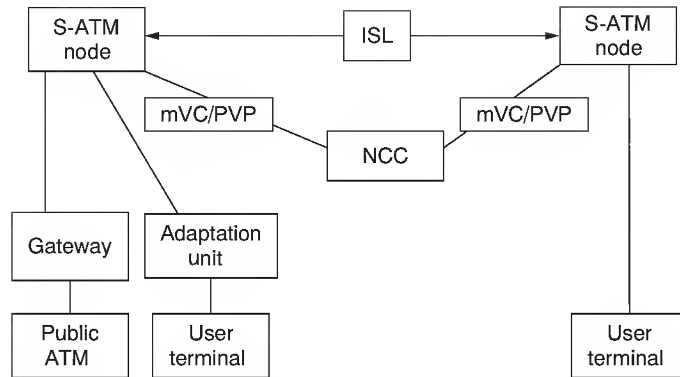
LEO satellites offer two main advantages over GEO satellites for networking. Because of the shorter ranges involved, the propagation delay is very much less, and, much lower transmit power is needed. Constellations

TABLE 15.3 Satellite ATM Architectures

Network	Bent pipe	On board processing (OBP)
	Relay	
	Point to point linkage between two fixed ATM users.	Not applicable.
	Access	
Fixed ATM	UNI at user terminal; UNI/NNI at GES. No mobility support.	Media access required. UNI between user and satellite. No mobility support.
Mobile ATM	Mobility enhanced UNl at user terminal and NNl between GES and ATM network.	Mobility support provided by the mobility enhanced switching at terrestrial ATM and at satellite.
	Interconnect	
Fixed ATM	High speed interconnection between fixed ATM networks PNNl, B-lCI, or public UNl between GESs and ATM networks. No mobility support.	ATM switch on board satellite acts as intermediate node. Supports NNI signaling, cell switching, and multiplexing. No mobility support if GEOs used.
Mobile ATM	High speed interconnections between mobile and fixed ATM networks and between two mobile ATM networks. Mobility enhanced NNI between GESs and networks.	High speed interconnections between mobile and fixed ATM networks, and between two mobile ATM networks. Mobility enhanced NNI between GESs and networks.
	Full mesh	
SATM	Not applicable.	Satellites form an ATM network in space, which supports all of the above scenarios.

NOTES: ATM—Asynchronous Transfer Mode; B-ICI—Broadband Inter-Carrier Interface; GES—Gateway Earth Station; NNI—Network Node Interface; PNNI—Private Network Node Interface; SATM—Satellite ATM (network); UNI—User Network Interface.

of LEO satellites with on-board processing are employed. However, because the LEO satellites are not geostationary, antenna beam switching is required as the spot beam pattern sweeps across a given earth location. This is referred to as *intra-satellite switching*. Switching between *inter-satellite links* (ISLs) is also required as any given satellite moves out of the range of a particular earth location. This all adds to the complexity of the on-board requirements. Figure 15.8*a* is a block schematic of a typical ATM LEO satellite network (Todorova, 2002). User terminals may access the LEO satellite system directly, or they may require an *adaptation unit* (similar to the modem shown in Fig. 15.6). The public ATM network accesses the LEO satellite system through *gateways*



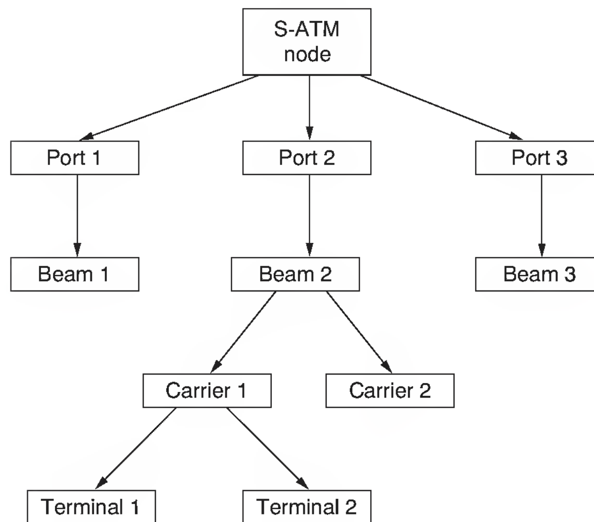


Figure 15.8 (*a*) Configuration of a network management system; (*b*) relationship between port, beam, carrier, and terminal. (*Courtesy of Petia Todovora.*)

(as in Fig. 15.7). Decentralized network management is proposed in the system described by Todorova, (2002), where overall management is provided by the *network control center* (NCC) but certain functions are carried out by the satellite ATM switches (S-ATM). The management information is carried in signaling channels termed *management virtual channels* (mVCs). For added security a stand-by path, in the form of a *permanent virtual path* (PVP), is provided in the event that the signaling channel should fail.

Figure 15.8*b* is a block schematic showing the relationship between port, beam, carrier, and terminal (Todorova and Nguyen, 2001). As shown, the ports from the ATM switch connect to individual beams. The beams may be multicarrier, two being shown in Fig. 15.8, and each carrier can be received by more than one earth terminal.

15.6 The Internet

On October 24, 1995, the *Federal Networking Council* (FNC) in the United States passed a resolution defining the Internet as a global information system that

1. Is logically linked together by a globally unique address space based on the *Internet protocol* (IP) or its subsequent extensions/follow-ons
2. Is able to support communications using the *transmission control protocol/Internet protocol* (TCP/IP) suite or its subsequent extensions/follow-ons, and/or other IP-compatible protocols
3. Provides, uses or makes accessible, either publicly or privately, high-level services layered on the communications and related infrastructure described herein.

This formal description of the Internet summarizes what in fact was many years of evolutionary growth and change (see Leiner et al., 2000). The key elements in this definition are the TCP and the IP, both of which are described shortly. These protocols are usually lumped together as TCP/IP and are embedded in the software for operating systems and browsers such as Windows and Netscape.

The Internet does not have its own physical structure. It makes use of existing physical plant, the copper wires, optical fibers, and satellite links, owned by companies such as AT&T, MCI, and Sprint. Although there is no identifiable structure, access to the Internet follows well-defined rules. Users connect to *Internet service providers* (ISPs), who in turn connect to *network service providers* (NSPs), who complete the connections to other users and to *servers*. Servers are computers dedicated to the purpose of providing information to the Internet. They run specialized software for each type of Internet application. These include email, discussion groups, long-distance computing, and file transfers. *Routers* are computers that form part of the communications net and that route or direct the data along the best available paths in the network.

Although there is no central management or authority for the Internet, its extraordinarily rapid growth has meant that some control has to be exercised over what is permitted. A summary of the controlling groups is shown in Fig. 15.9*a*. A description of the groups will be found in Leiner et al. (2000) and Mackenzie (1998).

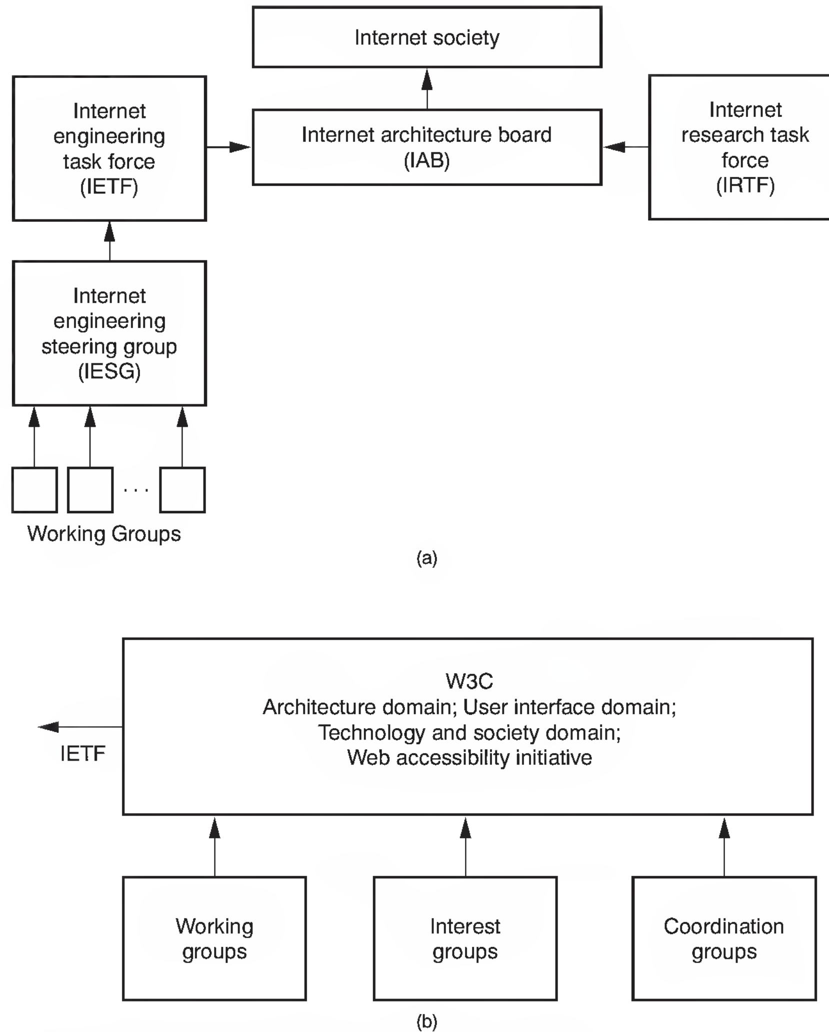


Figure 15.9 (*a*) Internet groups; (*b*) World Wide Web groups.

The *World Wide Web* (WWW) is probably the most widely used application on the Internet. The evolution and growth of the WWW has been rather similar to that of the Internet itself, with no central authority but still with a structure that attempts to regulate what happens. The WWW Consortium, referred to as W3C, was founded in October 1994 (Jacobs, 2000). W3C oversees a number of special interest groups, as shown in Fig. 15.9*b*, and coordinates its efforts with the IETF and with other standards bodies. Details of the W3C will be found in Jacobs (2000).

15.7 Internet Layers

The uplink and downlink between satellite and earth stations forms the *physical layer* in a data communication system. By *data communications* is meant communications between computers and peripheral equipment. The signals are digital, and although digital signals are covered in Chap. 10, the satellite links must be able to accommodate the special requirements imposed by networks. The terminology used in networks is highly specialized, and some of these terms are explained here to provide the background needed to understand the satellite aspects (see also Sec. 15.3). The Internet, of course, is a data communication system (although there is presently a move to incorporate voice communications along with data in what is known as *voice over Internet Protocol*, or *VoIP*).

The data are transmitted in *packets*. Many separate functions have to be performed in packet transmission, such as packet addressing, routing, and coping with packet congestion. The modern approach is to assign each function to a layer in what is termed the *network architecture*. This has already been encountered in connection with ATM, as shown in Fig. 15.1. The layers are conceptual in the sense that they may consist of software or some combination of software and hardware. In the case of the Internet, the network architecture is referred to the TCP/IP model, although there are protocols other than TCP/IP contained in the model. The layered structure is shown in Fig. 15.10. A brief description of these layers is included to familiarize the reader with some of the terms used in network communications, although the TCP layer is of most interest in this chapter.

- *Physical layer.* This covers such items as the physical connectors, signal format, modulation, and the uplink and downlink in a satellite communications system.
- *Data-link layer.* The function of this layer is to organize the digital data into blocks as required by the physical layer. For example, if the physical layer uses ATM technology, as described in Sec. 15.3, the

Applications & Services	
TCP	UDP
IP	
Data Link	
Physical	

Figure 15.10 Layered structure for TCP/IP. (*Courtesy of Feit, 1997.*)

data are organized into cells. Digital transmission by satellite frequently uses TDMA, as described in Chap. 14, and satellite systems are being developed which transmit Internet data over ATM. Thus the data-link layer has to organize the data into a suitable format to suit the physical-layer technology. In the terrestrial Internet, the data link converts the data into frames. The data-link layer and the physical layer are closely interrelated, and it can be difficult sometimes to identify the interface between these two layers (Mackenzie, 1998).

- *Network layer.* This is strictly an IP layer. The packets are passed along the Internet from router to router and to the host stations. No exact path is laid out beforehand, and the IP layers in the routers must provide the destination address for the next leg of the journey so to speak. This destination address is part of the IP header attached to the packet. The source address is also included as part of the IP header. The problems of lost packets or packets arriving out of sequence are not a concern of the IP layer, and for this reason, the IP layer is called *connectionless* (i.e., it does not require a connection to be established before sending a packet on). These problems are taken care of by the transport layer.
- *Transport layer.* Two sets of protocol are provided in this layer. With the TCP, information is passed back and forth between transport layers, which controls the information flow. This includes such functions as the correct sequencing of packets, replacement of lost packets, and adjusting the transmission rate of packets to prevent congestion. In the early days of the Internet when traffic was comparatively light, these problems could be handled even where satellite transmissions were involved. With the enormous increase in traffic on the present-day Internet, these problems require special solutions where satellite systems are used, which are discussed in later sections. The TCP layer is termed *connection-oriented* (compared with the connectionless service mentioned above) because the sender and receiver must be in communication with each other to implement the protocol. There are situations where a simple standalone message may need to be sent which does not require the more complex TCP. For these types of message, another transport layer protocol called the *user datagram protocol* (UDP) is used. The UDP provides a connectionless service, similar to IP. The UDP header adds the port numbers for the source and destination applications.

The term *packet* has been used somewhat loosely up to this point. A more precise terminology is used for packets at the various layers, and this is shown in Fig. 15.11. At the application level the packet is simply

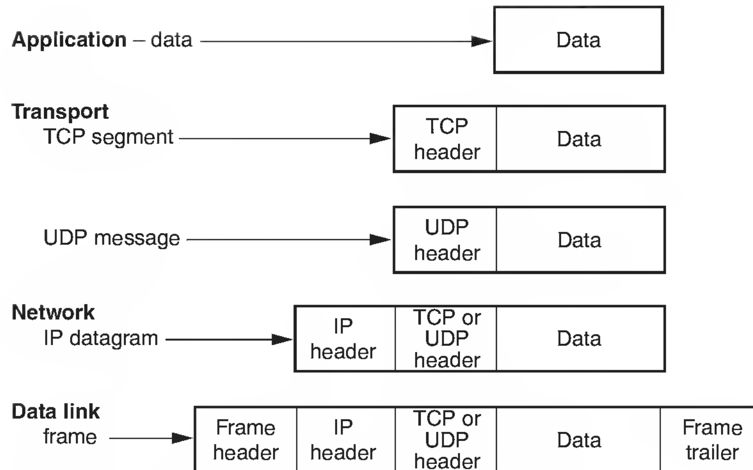


Figure 15.11 Packet terminology. (*Courtesy of Feit, 1997.*)

referred to as *data*. The packet comprising the TCP header, and the data are a *TCP segment*. The packet comprising the UDP header and the data is a *UDP message*. The packet comprising the IP header, the TCP or UDP header, and the data is an *IP datagram*. Finally, the packet comprising the data-link frame header, the frame trailer (used for error control), and the IP datagram is a frame.

It should be noted here that the preceding definitions are those used in version 4 of the IPv4. IPv6 is a more recent version being brought on-stream in which the IP datagram is in fact called an *IP packet*.

Some of the units used in data transmission are:

- *Byte.* Common usage has established the byte (symbol B) as a unit of 8 bits, and this practice will be followed here. It should be noted, however, that in computer terminology, a byte can mean a unit other than 8 bits, and the 8-bit unit may be called an *octet*.
- *Kilobyte.* The kilobyte (symbol kB) is 1024 bytes. Transmission rates may be stated in kilobytes per second or kB/s.
- *Megabyte.* The megabyte (symbol MB) is 1024 kilobytes. Transmission rates may be stated in megabytes per second or MB/s.

The TCP/IP suite is shown in Fig. 15.12, and an excellent detailed description of these protocols will be found in Feit (1997). The present text will be concerned more with the special enhancements needed on TCP/IP for successful satellite transmission.

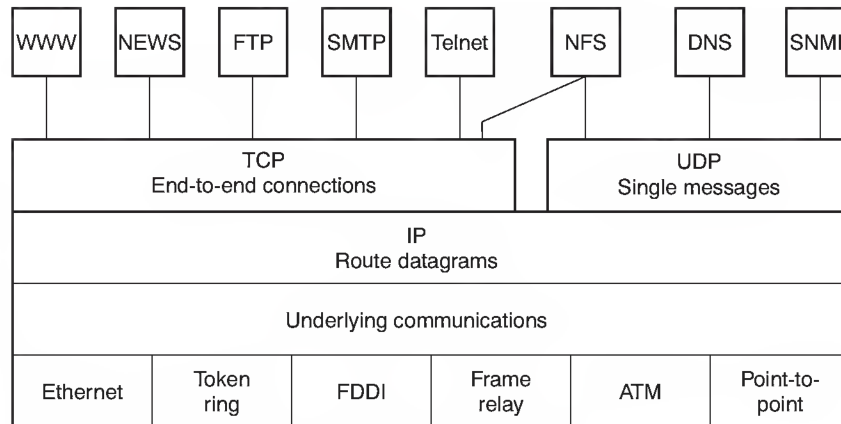


Figure 15.12 The TCP/IP suite. (*Courtesy of Feit, 1997.*)

15.8 The TCP Link

A *virtual communications link* exists between corresponding layers in a network. The header in the TCP segment (see Fig. 15.11) carries instructions that enable communication between the send and receive TCP layers. Of course, the communication has to pass through the other layers and along the physical link, but only the TCP layers act on the TCPs contained in the segment header. There is no direct physical link between the TCP layers, and for this reason, it is called a *virtual link*.

The send and receive TCP layers have buffer memories (usually just called *buffers*). The receive buffer holds incoming data while they are being processed. The send buffer holds data until they are ready for transmission. It also holds copies of data already sent until it receives an acknowledgment that the original has been received correctly. The *receive window* is the amount of receive buffer space available at any given time. This changes as the received data are processed and removed from the buffer. The receive TCP layer sends an *acknowledgment* (ACK) signal to the send TCP layer when it has cleared data from its buffer, and the ACK signal also provides an update on the current size of the receive window.

The send TCP layer keeps track of the amount of data in transit and, therefore, unacknowledged. It can calculate the amount of receive buffer space remaining, allowing for the data in transit. This remaining buffer space represents the amount of data that can still be sent and is termed the *send window*. The send TCP layer also sets a *timeout period*, and failure to receive an ACK signal within this period results in a duplicate packet being sent. On terrestrial networks, the probability of bit error (see Chap. 10) is extremely low, and *congestion* is the most likely

reason for loss of ACK signals. Because a network carries traffic from many sources, traffic congestion can occur. The IP layer of the TCP/IP discards packets when congestion occurs, and hence the corresponding ACK signals from the TCP layer do not get sent. Rather than continually resending packets, the send station reduces its rate of transmission, this being known as *congestion control*. A *congestion window* is applied, which starts at a size of one segment for a new connection. The window is doubled in size for each ACK received until it reaches a maximum value determined by the number of failed ACKs experienced. For normal operation, the congestion window grows in size to equal the receive window. The congestion window increases slowly at first, but as each doubling takes effect, the size increases exponentially. This controlling mechanism is known as *slow start*. If congestion sets in, this will be evidenced by an increase in the failure to receive ACKs, and the send TCP will revert to the slow start.

15.9 Satellite Links and TCP

Although satellite links have formed part of the Internet from its beginning, the rapid expansion of the Internet and the need to introduce congestion control have highlighted certain performance limitations imposed by the satellite links. Before discussing these, it should be pointed out that the increasing demand for Internet services may well be met best with satellite direct-to-home links, and many companies are actively engaged in setting up just such systems. In the ideal case, the virtual link between TCP layers should not be affected by the physical link, and certainly the TCP is so well established that it would be undesirable (some would say unacceptable) to modify it to accommodate peculiarities of the physical link. The factors that can adversely affect TCP performance over satellite links are as follows:

Bit error rate (BER). Satellite links have a higher bit error rate (BER; see Chap. 10) than the terrestrial links forming the Internet. Typically, the satellite link BER without error-control coding is around 10^{-6}, whereas a level of 10^{-8} or lower is needed for successful TCP transfer (Chotikapong and Sun, 2000). The comparatively low BER on terrestrial links means that most packet losses are the result of congestion, and the TCP send layer is programmed to act on this assumption. When packets are lost as a result of high BER, therefore, as they might on satellite links, the TCP layer assumes that congestion is at fault and automatically invokes the congestion control measures. This slows the throughput.

Round-trip time (RTT). The round-trip time (RTT) of interest here is the time interval that elapses between sending a TCP segment and

receiving its ACK. With geostationary (GEO) satellites, the round-trip propagation path is ground station-to-satellite-to-ground station and back again. The range from ground station to the satellite (see Chap. 3) is on the order of 40,000 km, and therefore, the propagation path for the round trip is $4 \times 40{,}000 = 160{,}000$ km. The propagation delay is therefore $160{,}000 \times 10^3/(3 \times 10^8) = 0.532$ s. This is just the space propagation delay. The total round-trip time must take into account the propagation delays on the terrestrial circuits and the delays resulting from signal processing. For order of magnitude calculations, an RTT value of 0.55 s would be appropriate. The send TCP layer must wait this length of time to receive the ACKs, and of course, it cannot send new segments until the ACKs are received, which is going to slow the throughput. The send TCP timeout period is also based on the RTT, and this will be unduly lengthened. Also, with interactive applications, such as Telnet, this delay is highly undesirable.

Bandwidth-delay product (BDP). The RTT is also used in determining an important factor known as the *bandwidth delay product* (BDP). The delay part of this refers to the RTT, since a sender has to wait this amount of time for the ACK before sending more data. The bandwidth refers to the channel bandwidth. As shown in Chaps. 10 and 12, bandwidth and bit rate are directly related. In network terminology, the bandwidth is usually specified in bytes per second (or multiples of this), where it is understood that 1 byte is equal to 8 bits. For example, a satellite bandwidth of 36 MHz carrying a BPSK signal could handle a bit rate given by Eq. (14.30) as 30 Mb/s. This is equivalent to 3.75×10^6 B/s or about 3662 kB/s. If the sender transmits at this rate, the largest packet it can send within the RTT of 0.55 s is $3662 \times 0.55 = 2014$ kB approximately. This is the BDP for the two-way satellite channel. The channel is sometimes referred to as a pipeline, and one that has a high BDP, as a long fat pipe. Now the receive TCP layer uses a 16-bit word to notify the send TCP layer of the size of the receive window it is going to use. Allowing 1 byte for certain overheads, the biggest segment size that can be declared for the receive window is $2^{16} - 1 = 65{,}535$ bytes, or approximately 64 kB. (Recall that 1 kB is equal to 1024 bytes.) This falls well short of the 2014 kilobytes set by the BDP for the channel, and thus the channel is very underutilized.

Variable round-trip time. Where lower earth orbiting satellites are used such as those in LEOs and MEOs, the propagation delays will be much less than that for the GEO. The slant range to LEOs is typically on the order of a few thousand kilometers at most, and for MEOs, a few tens of thousand kilometers. The problem with these orbits is not so much the absolute value of delay as the variability. Because these satellites are not geostationary, the slant range varies,

and for continuous communications there is the need for intersatellite links, which also adds to the delay and the variability. For example, for LEOs, the delay can vary from a few to about 80 ms. Whether or not this will have an impact on TCP performance is currently an open question (RFC-2488).

15.10 Enhancing TCP Over Satellite Channels Using Standard Mechanisms (RFC-2488)

In keeping with the objective that, where possible, the TCP itself should not be modified to accommodate satellite links, the *Request for Comments 2488* (RFC-2488) describes in detail several ways in which the performance over satellite links can be improved. These are summarized in Table 15.4. The first two mechanisms listed do not require any changes to the TCP. The others do require extensions to the TCP. As always, any extensions to the TCP must maintain compatibility with networks that do not employ the extensions. Brief descriptions of the mechanisms are included, but the reader is referred to the *Requests for Comments* (RFCs) for full details (see Sec. 15.11).

MTU stands for *maximum transmission unit*, and *Path MTU-Discovery* is a method that allows the sender to find the largest packet and, hence, largest TCP segment size that can be sent without fragmentation. The congestion window is incremented in segments; hence, larger segments allow the congestion window to increment faster in terms of number of bytes carried. There is a delay involved in implementing Path MTU-Discovery, and of course, there is the added complexity. Overall, however, it improves the performance of TCP over satellite links.

TABLE 15.4 Summary of Objectives in RFC-2488

Mechanism	Use	RFC-2488 section	Where applied
Path MTU-Discovery	Recommended	3.1	Sender
FEC	Recommended	3.2	Link
	TCP congestion control		
Slow start	Required	4.1.1	Sender
Congestion avoidance	Required	4.1.1	Sender
Fast retransmit	Recommended	4.1.2	Sender
Fast recovery	Recommended	4.1.2	Sender
	TCP large windows		
Window scaling	Recommended	4.2	Sender and receiver
PAWS	Recommended	4.2	Sender and receiver
RTTM	Recommended	4.2	Sender and receiver
TCP SACKS	Recommended	4.4	Sender and receiver

Forward error correction (FEC). *Lost packets, whether from transmission errors or* congestion, are assumed by the TCP to happen as a result of congestion, which means that congestion control is implemented, with its resulting reduction in throughput. Although there is ongoing research into ways of identifying the mechanisms for packet loss, the problem still remains. Application of FEC (as described in Chap. 11) therefore should be used where possible.

Slow start and congestion avoidance. These strategies have already been described in Sec. 15.8, along with the problems introduced by long RTTs. Slow start and congestion avoidance control the number of segments transmitted, but not the size of the segments. Using Path MTU-Discovery as described earlier can increase the size, and hence the data throughput is improved.

Fast retransmit and fast recovery. From the nature of the ACKs received, the fast retransmit algorithm enables the sender to identify and resend a lost segment before its timeout expires. Since the data flow is not interrupted by timeouts, the sender can infer that congestion is not a problem, and the fast recovery algorithm prevents the congestion window from reverting to slow start. The fast retransmit algorithm can only respond to one lost segment per send window. If there is more than one, the others trigger the slow start mechanism.

TCP large windows. As shown in Sec. 15.9 in connection with the bandwidth delay product, the receive window size is limited by the address field to 64 kilobytes maximum. By introducing a window scale extension into the TCP header, the address field can be effectively increased to 32 bits. Allowing for certain overheads, the maximum window size that can be declared is 2^{30} = 1 gigabyte (again keeping in mind that 1 gigabyte = 1024^3 bytes). The window size and hence the scale factor can be set locally by the receive TCP layer. Note, however, that the TCP extension has to be implemented at the sender and the receiver.

The two mechanisms PAWS, which stands for *protection against wrapped sequence*, and RTTM, which stands for *round-trip time measurement*, are extensions that should be used with large windows. Maintaining steady traffic flow and avoiding congestion require a current knowledge of the RTT, which can be difficult to obtain with large windows. By including a *time stamp* in the TCP header, the RTT can be measured. Another problem that arises with large windows is that the numbering of old sequences can overlap with new, a condition known as *wrap-around*. The protection against wrapped sequences is an algorithm that also makes use of the time stamp. These algorithms are described fully in RFC-1323.

SACK stands for *selective acknowledgment* and is a strategy that enables the receiver to inform the sender of all segments received successfully. The sender then need resend only the missing segments. The strategy should be used where multiple segments may be lost during transmission, such as, for example, in a satellite link, since clearly, retransmission of duplicate segments over long delay paths would seriously reduce the throughput. Full details of SACK will be found in RFC-2018.

15.11 Requests for Comments

The rapid growth of the Internet resulted, in large part, from the free and open access to documentation provided by network researchers. The ideas and proposals of researchers are circulated in memos called *requests for comments* (RFCs). They can be accessed on the World Wide Web at a number of sites, for example, http://www.rfceditor.org/. Following is a summary of some of the RFCs that relate specifically to satellite links and have been referred to in Sec. 15.5

- *RFC-2760, Ongoing TCP Research Related to Satellites, February 2000. Abstract*: This document outlines possible TCP enhancements that may allow TCP to better utilize the available bandwidth provided by networks containing satellite links. The algorithms and mechanisms outlined have not been judged to be mature enough to be recommended by the IETF. The goal of this document is to educate researchers as to the current work and progress being done in TCP research related to satellite networks.
- *RFC-2488, Enhancing TCP Over Satellite Channels Using Standard Mechanisms, January 1999. Abstract*: The TCP provides reliable delivery of data across any network path, including network paths containing satellite channels. While TCP works over satellite channels, there are several IETF standardized mechanisms that enable TCP to more effectively utilize the available capacity of the network path. This document outlines some of these TCP mitigations. At this time, all mitigations discussed in this document are IETF standards track mechanisms (or are compliant with IETF standards).
- *RFC-2018, TCP Selective Acknowledgment Options, October 1996. Abstract*: TCP may experience poor performance when multiple packets are lost from one window of data. With the limited information available from cumulative acknowledgments, a TCP sender can only learn about a single lost packet per round-trip time. An aggressive sender could choose to retransmit packets early, but such retransmitted segments may have already been received successfully. A SACK

mechanism, combined with a selective repeat retransmission policy, can help to overcome these limitations. The receiving TCP sends back SACK packets to the sender informing the sender of data that have been received. The sender can then retransmit only the missing data segments. This memo proposes an implementation of SACK and discusses its performance and related issues.

- *RFC-1323, TCP Extensions for High Performance, May 1992. Abstract*: This memo presents a set of TCP extensions to improve performance over large bandwidth delay product paths and to provide reliable operation over very high-speed paths. It defines new TCP options for scaled windows and timestamps, which are designed to provide compatible interworking with TCPs that do not implement the extensions. The timestamps are used for two distinct mechanisms: RTTM and PAWS. Selective acknowledgments are not included in this memo. This memo combines and supersedes RFC-1072 and RFC-1185, adding additional clarification and more detailed specification. App. C of RFC-1323 summarizes the changes from the earlier RFCs.
- *RFC-1072, TCP Extensions for Long Delay Paths, October 1988. Status of this memo*: This memo proposes a set of extensions to the TCP to provide efficient operation over a path with a high band-width delay product. These extensions are not proposed as an Internet standard at this time. Instead, they are intended as a basis for further experimentation and research on TCP performance. Distribution of this memo is unlimited.

15.12 Split TCP Connections

The TCP provides end-to-end connection. This means that the TCP layers at the sender and receiver are connected through a virtual link (see Sec. 15.3) so that such matters as congestion control and regulation of data flow, can be carried out without intervention of intermediate stages. It is to preserve this end-to-end connection that many of the extensions to TCP described in the preceding section have been introduced.

If, however, it is assumed that the end-to-end connectivity can be split, new possibilities are opened up for the introduction of satellite links as part of the overall Internet. Figure 15.13 shows one possible arrangement (Ghani and Dixit, 1999). Breaking the network in this way is termed *spoofing*. This refers to the fact that the TCP source thinks it is connected to the TCP destination, whereas the *interworking unit* (IWU) performs a protocol conversion. In Fig. 15.13, TCP Reno refers to the TCP with extensions: slow start, congestion avoidance, fast retransmit, fast

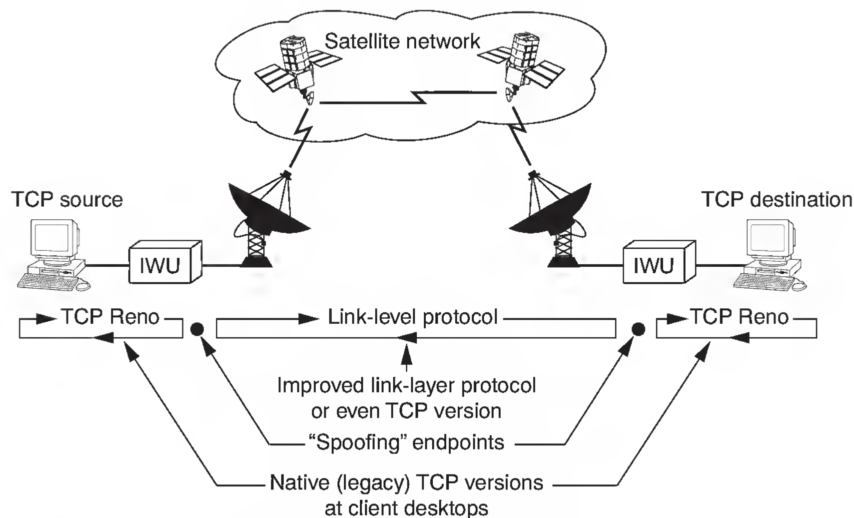


Figure 15.13 TCP/IP satellite link spoofing configuration. (*Courtesy of Ghani and Dixit, 1999. Copyright, 1999 IEEE.*)

recovery, support for large windows, and delayed ACKs. At the IWU the data are transferred from the TCP Reno to the data-link protocol. As shown in the figure, any one of a number of link layer protocols may be used. At the destination end, the IWU performs the conversion back to TCP Reno.

One approach developed at Roke Manor Research, Ltd. (West and McCann, 2000) illustrates some of the possibilities and problems associated with splitting. The two key issues are setup and teardown (West and McCann, 2000). To illustrate the process, consider a connection being set up between host A and host B. In setting up the connection, host A sends a synchronizing segment, labeled SYN in the TCP/IP scheme, which specifies certain protocols to be followed. Host B responds with its own SYN, which contains its protocol requirements, also an ACK signal that carries the number to be used by host A for the first data byte it sends. Host A then responds with an ACK signal that carries the number to be used by host B for the first data byte it sends. The three signals, SYN, SYN/ACK, and ACK, constitute, what is known as a *three-way handshake*.

A three-way handshake is also used to close (teardown). Suppose host B wishes to close. It sends a *final segment* (FIN). Host A acknowledges the FIN with an ACK and follows this with its own FIN. Host B acknowledges the FIN with an ACK. On connections with long RTTs, it will be seen that these three-way handshakes will be very time-consuming.

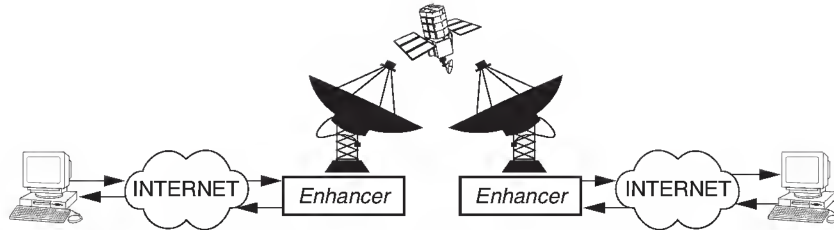


Figure 15.14 Physical architecture incorporating enhancer technology. (*Courtesy of West and McCann, 2000.*)

Figure 15.14 shows the system developed at Roke Manor. The *enhancers* perform the same function as the IWUs in Fig. 15.5 in that they terminate the Internet connections and do not require any modifications to the TCP/IP. A propriety protocol is used over the satellite link. Figure 15.15 illustrates a situation where both setup and teardown are spoofed. In this illustration, host B refuses the connection, but host A receives the RESET signal too late. The spoofed FIN ACK tells host A that the data transfer was successful.

Figure 15.16 shows a more appropriate strategy. In this case, the FIN sent by host A is not spoofed. Since host B has refused the connection, host A receives no FIN, ACK back from host B. Therefore, host A can infer from this that there was an error. While this is not as good as a regular TCP/IP connection, which would have reported a failure to connect on the first SYN, the system does adjust to the error and removes the spoofing on the setup on a second try.

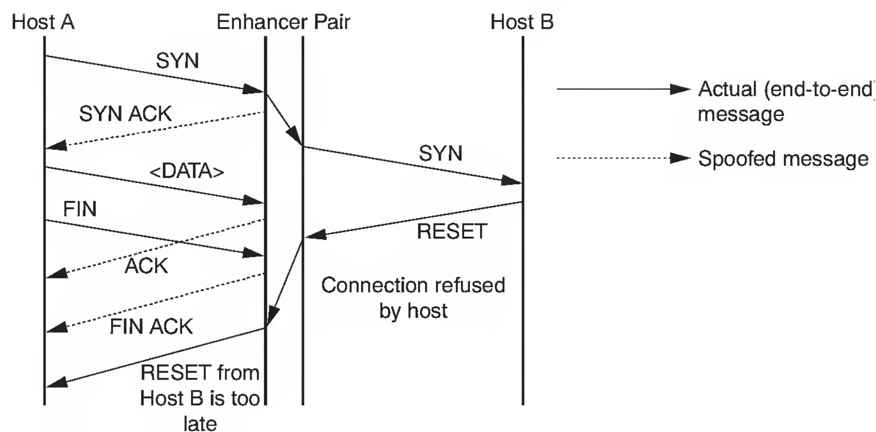


Figure 15.15 Connection establishment and closing. (*Courtesy of West and McCann, 2000.*)

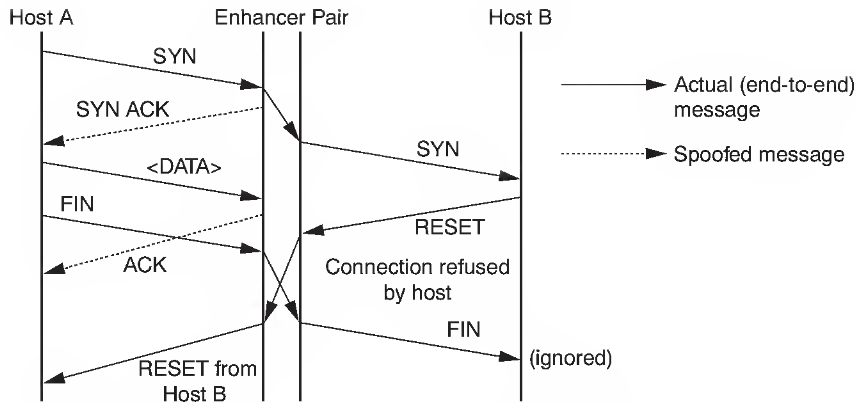


Figure 15.16 Improved connection close. (*Courtesy of West and McCann, 2000.*)

15.13 Asymmetric Channels

The term *asymmetry* applies in two senses to an Internet connection. It can refer to the data flow, which is often asymmetric in nature. A short request being sent for a Web page and the returned Web page may be a much larger document. Also, the acknowledgment packets sent on the return or reverse link are generally shorter than the TCP segments sent on the forward link. Values of 1500 bytes for data segments on the forward link and 40 bytes for ACKs on the reverse link are given in RFC-2760.

Asymmetry is also used to describe the physical capacities of the links. For small earth stations (e.g., VSATs), transmit power and antenna size (in effect, the EIRP) limit the uplink data rate, which therefore may be much less than the downlink data rate. Such asymmetry can result in ACK congestion. Again, using some values given in RFC-2760, for a 1.5 Mb/s data link a reverse link of less than 20 kb/s can result in ACK congestion. The levels of asymmetry that lead to ACK congestion are readily encountered in VSAT networks that share the uplink through multiple access.

In some situations, the reverse link may be completed through a terrestrial circuit, as shown in Fig. 15.17 (Ghani and Dixit, 1999). Here, the TCP source is connected to the satellite uplink through an IWU as before. The downlink signal feeds the small residential receiver, which is a receive-only earth station. An IWU on the receive side converts the data to the TCP format and sends them on to the destination. The ACK packets from the TCP destination are returned to the TCP source through a terrestrial network. As pointed out in RFC-2760, the reverse link capacity is limited not only by its bandwidth but also by queue lengths at routers, which again can result in ACK congestion. Some of the proposed

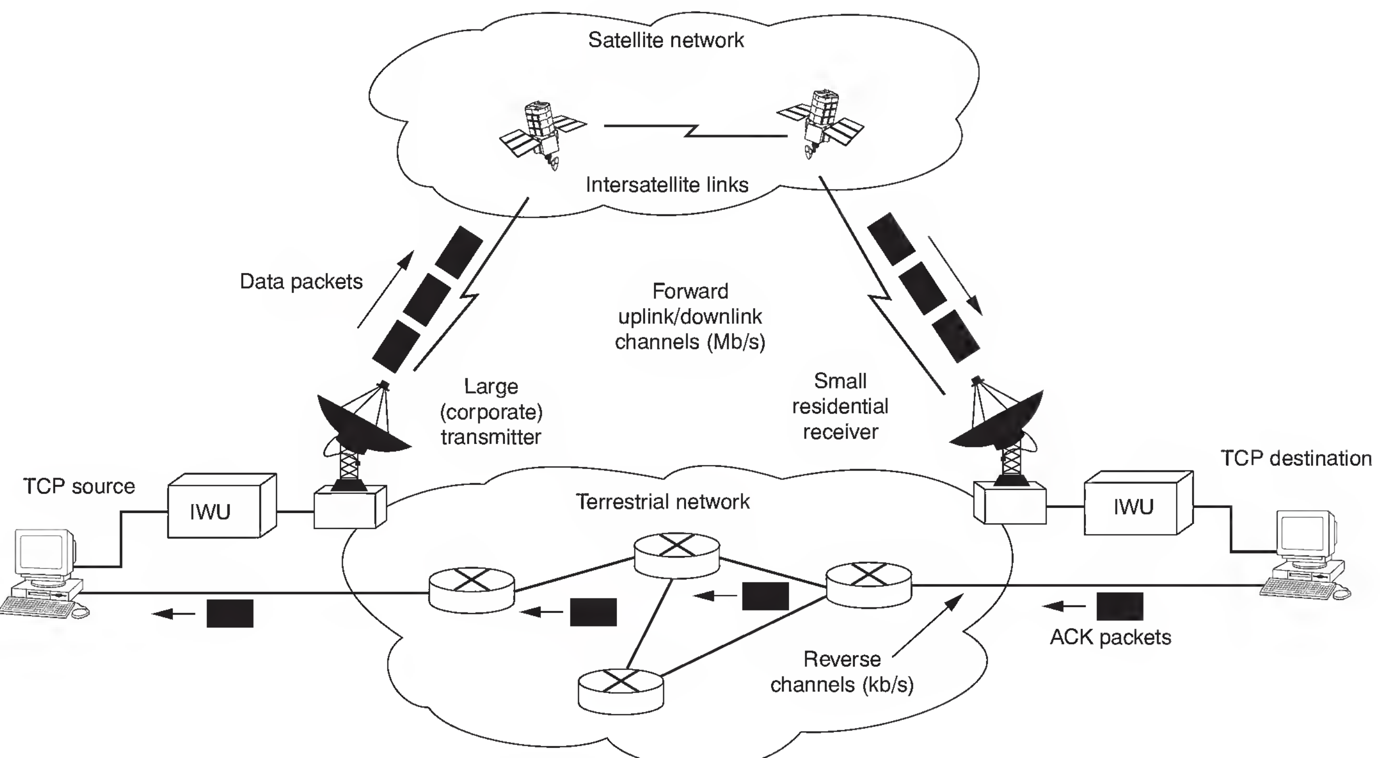


Figure 15.17 Asymmetric reverse ACK channel configuration. (*Courtesy of Ghani and Dixit, 1999. Copyright, 1999 IEEE.*)

methods of handling asymmetry problems and ongoing research are described in Ghani and Dixit (1999).

15.14 Proposed Systems

Most of the currently employed satellites operate in what is called a "bent pipe" mode, that is, they relay the data from one host to another without any onboard processing. Also, many of the problems with using geostationary satellites for Internet traffic arise because of the long propagation delay and the resulting high delay-bandwidth product. In some of the newer satellite systems, use is made of LEO and MEO satellites to cut down on the propagation delay time. Also, onboard signal processing is used in many instances that may result in the TCP/IP protocol being exchanged for propriety protocols over the satellite links. The satellite part of the network may carry the IP over ATM. The Ka band, which covers from 27 to 40 GHz, is used in many (but not all) of these newer systems. Wider bandwidths are available for carriers in the Ka band compared with those in the Ku band. A survey of some of these broadband systems will be found in Farserotu and Prasad (2000), but it should be pointed out that company changes in the form of mergers and takeovers occur that can drastically alter the services a company may offer. With this in mind, details of existing and proposed satellite internet services can be found using a search engine, such as Google. Some of the company names to search for are: Astrolink; Loral Cyberstar; Skybridge; Spaceway; iSky; and Teledesic.

15.15 Problems and Exercises

15.1. Write brief notes on the defining features of a broadband network.

15.2. Explain the difference between a wideband network and a broadband network. In network terminology, bandwidth is usually given in terms of bit rate, while in radio systems bandwidth is given in terms of frequency. Show how these concepts are connected.

15.3. Explain briefly what is meant by a *network protocol*. By referring to any of the standard publications showing the *open systems interconnection* (OSI) model, show how the ATM layer model, of Fig. 15.2, relates to the OSI model. (A comprehensive description of the OSI model can be found on the web site of Wikipedia, the free encyclopedia).

15.4. Describe briefly the difference between synchronous and asynchronous transfer modes. One author (Ramteke, 1994) points out that *synchronous transmission mode* (STM) uses asynchronous multiplexing, and ATM is transmitted using synchronous multiplexing (SONET). Explain how this is.

15.5. Describe the different types of interfaces and their functions in an ATM network.

15.6. With a BER of 10^{-8} the probability of a cell being discarded is about 10^{-13} and the probability of cells with undetected errors getting through is about 10^{-20} (Goralski, 1995, p. 140). On average, how many cells would be transmitted before a discard occurs? Is the probability of a cell with undetected errors getting through of any concern?

15.7. Describe briefly the difference between a ATM digital cross connect switch, and an ATM switch.

15.8. Describe briefly the difference between permanent virtual circuits and switched virtual circuits.

15.9. Explain what is meant by *bandwidth on demand* and how this is achieved in ATM.

15.10. A single user makes use of all the cells in an ATM transmission at a speed of 50 Mbps. What is the payload bandwidth?

15.11. An ATM system transmits 8000 frames per second with a maximum of 10 cells per frame. What is the speed of the link?

15.12. What type of signals are sensitive to variations in cell delay, and how does ATM accommodate these signals?

15.13. State the essential differences between the five classes of service offered with ATM, and the applications for which each class would be used. What is meant by quality of service?

15.14. What are the main technical parameters used in measuring ATM performance?

15.15. Discuss briefly the main problems encountered in incorporating satellite links in ATM networks, and the steps taken to overcome these.

15.16. Explain what is meant by a "bent pipe" system in connection with satellite communications.

15.17. Describe the main distinguishing features between satellite relay, satellite access, and satellite interconnect, in connection with ATM over satellite.

15.18. What are the main advantages and disadvantages in using LEO satellites compared with GEO satellites in ATM networks?

15.19. Write brief notes on the Internet and the WWW showing the relationship between them.

15.20. Describe the layered architecture that applies to networks, and indicate which of these layers contain (*a*) the IP and (*b*) the TCP. Explain what is meant by protocol.

15.21. Describe how *connectionless* and *connection-oriented* layers differ.

15.22. Describe what is meant by a *packet* in the different layers of a network.

15.23. Binary transmission occurs at a rate of 5000 bytes per second. What is this in bits per second?

15.24. Binary transmission occurs at a rate of 25 kilobytes per second. What is this in kilobits per second?

15.25. Binary transmission occurs at a rate of 30 megabytes per second. What is this in megabits per second?

15.26. What is meant by a *virtual link* in a network?

15.27. Define and explain what is meant by (*a*) *receive window,* (*b*) *send window,* and (*c*) *timeout* with reference to the Internet.

15.28. Explain what is meant by *congestion* and *slow start* in relation to Internet traffic.

15.29. In the early days of the Internet, Internet traffic could be sent over satellite links without any particular difficulty. State briefly the changes that have taken place that introduce difficulties and that require special attention.

15.30. Explain what is meant by RTT. Given that the uplink range to a satellite is 38,000 km and the downlink range is 40,000 km, calculate the RTT.

15.31. The overall bandwidth for a satellite link is 36 MHz, and the filtering is raised-cosine with a rolloff factor of 0.2. BPSK modulation is used. The range to the satellite is 39,000 km in both directions. Calculate the size of the largest packet that could be in transit if the BDP was the limiting factor. Assume that the RTT is equal to the space propagation delay.

15.32. A LEO satellite is used for Internet transmissions. The orbit can be assumed circular at an altitude of 800 km. Assuming that the minimum usable angle of elevation of the satellite is 10° (the angle above the horizon), calculate the approximate values of maximum and minimum RTTs. A spherical earth of uniform mass and radius of 6371 km may be assumed.

15.33. Repeat Prob. 15.32 for a circular MEO at altitude 20,000 km.

15.34. Briefly discuss the enhancements covered in RFC-2488 and how these might improve the performance of Internet traffic over satellite links.

15.35. By accessing the RFC site on the WWW, find the latest version of RFC-2760. Comment on this.

15.36. Explain what is meant by *split TCP connections* and why these might be considered undesirable for Internet use.

15.37. Explain what is meant by *asymmetric channels*. Describe how asymmetric channels may be incorporated in Internet connections via satellites.

References

Akyildiz, I. F., I. Joe, H. Driver, and Y. L. Ho. 1998. "A New Adaptive FEC Scheme for Wireless Networks." *Proc. IEEE Milcom'98*, Boston, MA, October.

Chotikapong, Y., and Z. Sun. 2000. "Evaluation of Application Performance for TCP/IP via Satellite Links." *IEE (London) Aerospace Group Seminar: Satellite Services and the Internet*, London. 17 February.

Ericsson. 2003. CLA-2000™/ATM Link Accelerator™, at http://www.componedex.com/products/cla_atm.htm

Farserotu, J., and R. Prasad. 2000. "A Survey of Future Broadband Multimedia Satellite Systems, Issues and Trends." *IEEE Commun. Mag.*, Vol. 38, No. 6, pp. 128–133, June.

Feit, S. 1997. *TCP/IP*, 2d ed. McGraw-Hill, New York.

Ghani, N., and S. Dixit. 1999. "TCP/IP Enhancements for Satellite Networks." *IEEE Commun. Mag.,* Vol. 37, No. 7, pp. 64–72, July. Goralski, W. J. 1995. *Introduction to ATM Networking*. McGraw-Hill, New York.

Hayt, W. H., and J. E. Kemmerly 1978. Engineering Circuit Analysis. McGraw-Hill, New York.

Ivancic, W. D., B. D. Frantz and M. J. Spells. 1998. "MPEG-2 Over Asynchronous Transfer Mode (ATM) Over Satellite Quality of Service (QoS) Experiments: Laboratory Tests." *NASA/TM/1998 206535*, September.

Ivancic, W. D., D. E. Brooks and B. D. Frantz. 1997. "ATM Quality of Service Tests for Digitized Video Using ATM Over Satellite: Laboratory Tests." *NASA/TM/107421*, July.

Jacobs, I. 2000. W3C. At http://www.w3.org/Consortium/, March.

Leiner, B. M., V. G. Cerf, D. D. Clark, R. E. Khan, L. Kleinrock, D. C. Lynch, J. Postel, L. G. Roberts, and S. Wolff. 2000. A Brief History of the Internet, at http://www.isoc.org/internet-history/brief.html, revised 14 April.

Mackenzie, L. 1998. *Communications and Networks.* McGraw-Hill, New York.

Ramteke, T. 1994. *Networks*. Prentice Hall, New Jersey.

Russell, T. 2000. *Telecommunications Protocols*. 2d ed., McGraw-Hill, New York.

Todorova, P. 2002. Network Management in ATM LEO Satellite Networks. Proceedings of the 35th Annual Hawaii International Conference on System Sciences. (Google search: Todorova network management ATM) Todorova, P., and H. N. Nguyen. 2001. "On-board Buffer Architectures for Low Earth (LEO) Satellite ATM Systems." *IEEE GLOBECOM 2001*, November 25–29, 2001, San Antonio, Texas.

Toh, C., and V. O. K. Li, 1998. "Satellite ATM Network Architectures: An Overview." *IEEE Network*, Vol. 12, No. 5, pp. 61–71, September/October. Walrand, J. 1991. *Communication Networks*. Aksen Associates, Homewood, IL.

Web ProForum Tutorials of the International Engineering Consortium, at http://www.lec.org.

West, M., and S. McCann. 2000. "Improved TCP Performance over Long-Delay and Error-Prone Links." *IEE (London) Aerospace Group Seminar, Satellite Services and the Internet*, London. 17 February.

Chapter

16

Direct Broadcast Satellite (DBS) Television

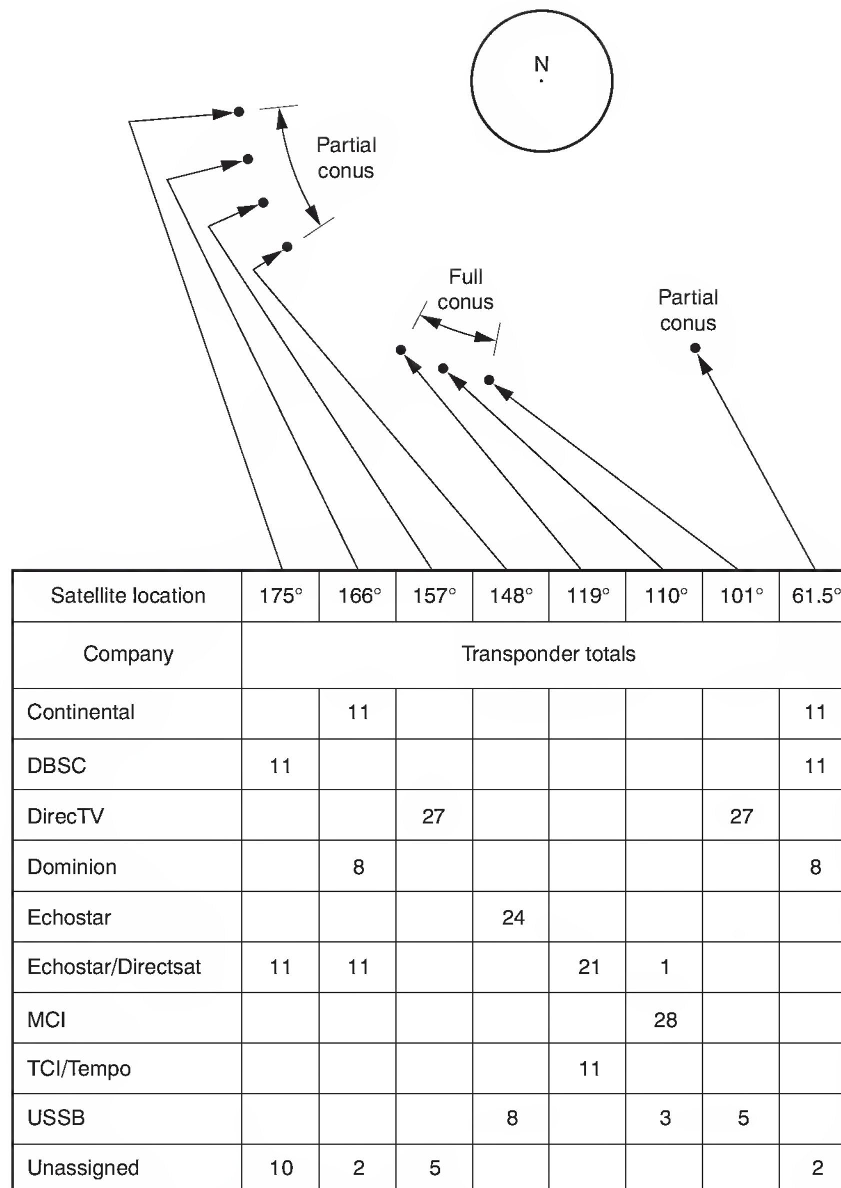
16.1 Introduction

Satellites provide *broadcast* transmissions in the fullest sense of the word, because antenna footprints can be made to cover large areas of the earth. The idea of using satellites to provide direct transmissions into the home has been around for many years, and the services provided are known generally as *direct broadcast satellite* (DBS) services. Broadcast services include audio, television, and Internet services. Direct broadcast television, which is digital TV, is the subject of this chapter.

A comprehensive overview covering the early years of DBS in Europe, the United States, and other countries is given in Prichard and Ogata (1990). Some of the regulatory and commercial aspects of European DBS will be found in Chaplin (1992), and the U.S. market is discussed in Reinhart (1990). Reinhart defines three categories of U.S. DBS systems, shown in Table 1.4. Of interest to the topic of this chapter is the high power category, the primary intended use of which is for DBS.

16.2 Orbital Spacing

From Table 1.4 it is seen that the orbital spacing is 9° for the high-power satellites, so adjacent satellite interference is considered nonexistent. The DBS orbital positions along with the transponder allocations for the United States are shown in Fig. 16.1. It should be noted that although the DBS services are spaced by 9°, *clusters of satellites* occupy



Satellite location	175°	166°	157°	148°	119°	110°	101°	61.5°
Company	Transponder totals							
Continental		11						11
DBSC	11							11
DirecTV			27				27	
Dominion		8						8
Echostar				24				
Echostar/Directsat	11	11			21	1		
MCI						28		
TCI/Tempo					11			
USSB				8		3	5	
Unassigned	10	2	5					2

Figure 16.1 DBS orbital positions for the United States.

the nominal orbital positions. For example, the following satellites are located at 119°W longitude: EchoStar VI, launched on July 14, 2000; EchoStar IV, launched on May 8, 1998; EchoStar II, launched September 10, 1996; and EchoStar I, launched on December 28,1995 (source: http://www.dishnetwork.com/).

16.3 Power Rating and Number of Transponders

From Table 1.4 it will be seen that satellites primarily intended for DBS have a higher [EIRP] than for the other categories, being in the range 51 to 60 dBW. At a *Regional Administrative Radio Council* (RARC) meeting in 1983, the value established for DBS was 57 dBW (Mead, 2000). Transponders are rated by the power output of their high-power amplifiers. Typically, a satellite may carry 32 transponders. If all 32 are in use, each will operate at the lower power rating of 120 W. By doubling up the high-power amplifiers, the number of transponders is reduced by half to 16, but each transponder operates at the higher power rating of 240 W. The power rating has a direct bearing on the bit rate that can be handled, as described in Sec. 16.8.

16.4 Frequencies and Polarization

The frequencies for direct broadcast satellites vary from region to region throughout the world, although these are generally in the Ku band. For region 2 (see Sec. 1.2), Table 1.4 shows that for high-power satellites, the primary use of which is for DBS, the uplink frequency range is 17.3 to 17.8 GHz, and the downlink range is 12.2 to 12.7 GHz. The medium-power satellites listed in Table 1.4 also operate in the Ku band at 14 to 14.5 GHz uplink and 11.7 to 12.2 GHz downlink. The primary use of these satellites, however, is for point-to-point applications, with an allowed additional use in the DBS service. In this chapter only the high-power satellites intended primarily for DBS will be discussed.

The available bandwidth (uplink and downlink) is seen to be 500 MHz. A total number of 32 transponder channels, each of bandwidth 24 MHz, can be accommodated. The bandwidth is sometimes specified as 27 MHz, but this includes a 3-MHz guardband allowance. Therefore, when calculating bit-rate capacity, the 24 MHz value is used, as shown in Sec. 16.5. The total of 32 transponders requires the use of both *right-hand circular polarization* (RHCP) and *left-hand circular polarization* (LHCP) in order to permit frequency reuse, and guard bands are inserted between channels of a given polarization. The DBS frequency plan for Region 2 is shown in Fig. 16.2.

16.5 Transponder Capacity

The 24-MHz bandwidth of a transponder is capable of carrying one analog television channel. To be commercially viable, DBS television [also known as *direct-to-home* (DTH) television] requires many more channels, and this requires a move from analog to digital television. Digitizing the audio and video components of a television program allows

RHCP

	1	3	5	. . .	31
Uplink MHz	17324.00	17353.16	17382.32		17761.40
Downlink MHz	12224.00	12253.16	12282.32		12661.40

LHCP

	2	4	6	. . .	32
Uplink MHz	17338.58	17367.74	17411.46		17775.98
Downlink MHz	12238.58	12267.74	12296.50		12675.98

Figure 16.2 The DBS frequency plan for Region 2.

signal compression to be applied, which greatly reduces the bandwidth required. The signal compression used in DBS is a highly complex process, and only a brief overview will be given here of the process. Before doing this, an estimate of the bit rate that can be carried in a 24-MHz transponder will be made.

From Eq. (10.16), the symbol rate that can be transmitted in a given bandwidth is

$$R_{\text{sym}} = \frac{B_{\text{IF}}}{1 + \rho} \tag{16.1}$$

Thus, with a bandwidth of 24 MHz and allowing for a rolloff factor of 0.2, the symbol rate is

$$R_{\text{sym}} = \frac{24 \times 10^6}{1 + 0.2} = 20 \times 10^6 \text{ symbols/s}$$

Satellite digital television uses QPSK. Thus, using $M = 4$ in Eq. (10.3) gives $m = 2$, and the bit rate from Eq. (10.5) is

$$R_b = 2 \times R_{\text{sym}} = 40 \text{ Mbps}$$

This is the bit rate that can be carried in the 24-MHz channel using QPSK.

16.6 Bit Rates for Digital Television

The bit rate for digital television depends very much on the picture format. One way of estimating the uncompressed bit rate is to multiply the number of pixels in a frame by the number of frames per second, and multiply this by the number of bits used to encode each pixel. The number of bits per pixel depends on the color depth per pixel, for example 16 bits

per pixel gives a color depth of $2^{16} = 65536$ colors. Using the HDTV format having a pixel count per frame of 1920×1080 and a refresh rate of 30 frames per second as shown in Table 16.1, the estimated bit rate is 995 Mbps. (A somewhat different estimate is sometimes used, which allows for 8 bits for each of the three primary colors, and this would result in a bit rate of approximately 1.49 Gbps for this version of HDTV).

From Table 16.1 it is seen that the uncompressed bit rate ranges from 118 Mb/s for standard definition television at the lowest pixel resolution to 995 Mb/s for high definition TV at the highest resolution. As a note of interest, the broadcast raster for studio-quality television, when digitized according to the international CCIR-601 television standard, requires a bit rate of 216 Mb/s (Netravali and Lippman, 1995).

A single DBS transponder has to carry somewhere between four and eight TV programs to be commercially viable (Mead, 2000). The programs may originate from a variety of sources, for example film, analog TV, and videocassette. Before transmission, these must all be converted to digital, compressed, and then *time-division multiplexed* (TDM). This TDM baseband signal is applied as QPSK modulation to the uplink carrier reaching a given transponder.

The compressed bit rate, and hence the number of channels that are carried, depends on the type of program material. Talk shows where

TABLE 16.1 ATSC Television Formats

No.	Format type	Name	Aspect ratio	Resolution pixels	Frames per second	Uncompressed bit rate, mbps
1	SDTV	480i	4:3	640 × 480	30	148
2	EDTV	480p	4:3	640 × 480	24	118
3	EDTV	480p	4:3	640 × 480	30	148
4	EDTV	480p	4:3	640 × 480	60	295
5	EDTV	480i	4:3	704 × 480	30	162
6	EDTV	480p	4:3	704 × 480	24	130
7	EDTV	480p	4:3	704 × 480	30	162
8	EDTV	480p	4:3	704 × 480	60	324
9	EDTV	480i	16:9	704 × 480	30	162
10	EDTV	480p	16:9	704 × 480	24	130
11	EDTV	480p	16:9	704 × 480	30	162
12	EDTV	480p	16:9	704 × 480	60	324
13	HDTV	720p	16:9	1280 × 720	24	334
14	HDTV	720p	16:9	1280 × 720	30	442
15	HDTV	720p	16:9	1280 × 720	60	885
16	HDTV	1080i	16:9	1920 × 1080	30	995
17	HDTV	1080p	16:9	1920 × 1080	24	796
18	HDTV	1080p	16:9	1920 × 1080	30	995

NOTES: ATSC—Advanced Television Systems Committee; HDTV—high-definition television; SDTV—standard definition television; EDTV—enhanced definition television; p—progressive scanning; i—interlaced scanning.

SOURCES: Booth, 1999; www.timefordvd.com, 2004.

there is little movement require the lowest bit rate, while sports channels with lots of movement require comparatively large bit rates. Typical values for SDTV are in the range of 4 Mb/s for a movie channel, 5 Mb/s for a variety channel, and 6 Mb/s for a sports channel (from MPEG and DSS Technical Notes (v0.3) by C. Fogg, 1995). Compression is carried out to *Moving Pictures Expert Group* (MPEG) standards.

16.7 MPEG Compression Standards

MPEG is a group within the *International Standards Organization and the International Electrochemical Commission* (ISO/IEC) that undertook the job of defining standards for the transmission and storage of moving pictures and sound. The standards are concerned only with the bit stream syntax and the decoding process, not with how encoding and decoding might be implemented. Syntax covers matters such as bit rate, picture resolution, time frames for audio, and the packet details for transmission. The design of hardware for the encoding and decoding processes is left to the equipment manufacturer. Comprehensive descriptions of the MPEG can be found at http://www.mpeg.org and in Sweet (1997) and Bhatt et al. (1997). The MPEG standards currently available are MPEG-1, MPEG-2, MPEG-4, and MPEG-7. For a brief explanation of the "missing numbers," see Koenen (1999).

In DBS systems, MPEG-2 is used for video compression. As a first or preprocessing step, the analog outputs from the red (R), green (G), and blue (B) color cameras are converted to a luminance component (Y) and two chrominance components (Cr) and (Cb). This is similar to the analog NTSC arrangement shown in Fig. 9.7, except that the coefficients of the matrix **M** (the 3 × 3 matrix) are different. In matrix notation, the equation relating the three primary colors to the Y, Cr, and Cb components is

$$\begin{bmatrix} \mathrm{Y} \\ \mathrm{Cr} \\ \mathrm{Cb} \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.168736 & -0.331264 & 0.5 \\ 0.5 & -0.418688 & -0.081312 \end{bmatrix} \begin{bmatrix} \mathrm{R} \\ \mathrm{G} \\ \mathrm{B} \end{bmatrix} \tag{16.2}$$

The Y, Cr, and Cb analog signals are sampled in the digitizer shown in Fig. 16.3. It is an observed fact that the human eye is less sensitive to resolution in the color components (Cr and Cb) than the luminance (Y) component. This allows a lower sampling rate to be used for the color components. This is referred to as *chroma subsampling*, and it represents one step in the compression process. MPEG-2 uses 4:2:0 sampling, which is described next.

Sampling is usually indicated by the ratios Y:U:V where Y represents the luminance (or luma) sampling rate, U the Cb sampling rate, and V the Cr sampling rate. The values for YUV are normalized to a value of

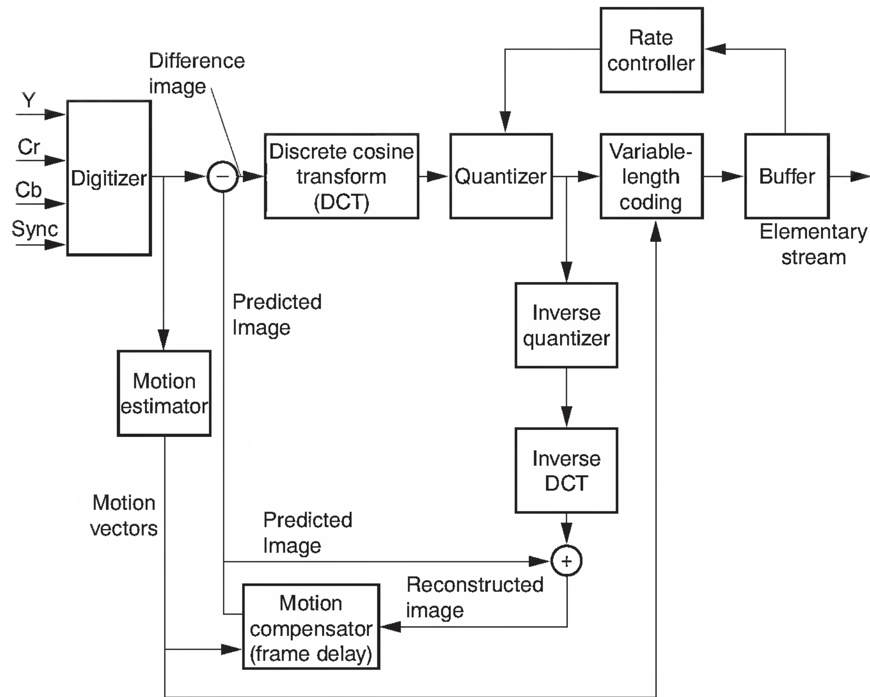


Figure 16.3 MPEG-2 encoder paths. (*Courtesy of Bhatt et al., 1997. IEEE.*)

4 for Y, and ratios commonly encountered with digital TV are 4:4:4, 4:2:2 and 4:2:0. These are explained next.

4:4:4 means that the sampling rates of Y, Cb, and Cr are equal. Each pixel would get three digital words, one for each of the component signals. If the words are 8-bits (commonly called a byte, but see Sec. 15.2) then each pixel would be encoded in 3 bytes.

4:2:2 means that the Cb and Cr signals are sampled at half the rate of the Y signal component. Every two pixels would have two bytes for the Y signal, one byte for the Cb signal and one byte for the Cr signal, resulting in 4 bytes for the 2-pixel block.

4:2:0 means that Cb and Cr are sampled at half the Y sampling rate, but they are sampled only on alternate scan lines. Thus vertical as well as horizontal resolution is reduced by half. A 2×2 pixel block would have 6 bytes, 4 bytes for Y, 1 byte for Cb and 1 byte for Cr.

Following the digitizer, difference signals are formed, and the *discrete cosine transform* (DCT) block converts these to a "spatial frequency" domain. The familiar Fourier transform transforms a time signal $g(t)$ to a frequency domain representation $G(f)$, allowing the signal to be filtered in the frequency domain. Here, the variables are time t and frequency f. In the DCT situation, the input signals are functions of the x

(horizontal) and y (vertical) space coordinates, $g(x, y)$. The DCT transforms these into a domain of new variables u and v, $G(u, v)$. The variables are called *spatial frequencies* in analogy with the time-frequency transform. It should be noted that $g(x, y)$ and $G(u, v)$ are *discrete functions*. In the quantizer following the DCT transform block, the discrete values of $G(u, v)$ are quantized to predetermined levels. This reduces the number of levels to be transmitted and therefore provides compression. The components of $G(u, v)$ at the higher spatial frequencies represent finer spatial resolution. The human eye is less sensitive to resolution at these high spatial frequencies; therefore, they can be quantized in much coarser steps. This results in further compression. (This step is analogous to the nonlinear quantization discussed in Sec. 10.3.)

Compression is also achieved through *motion estimation*. Frames in MPEG-2 are designated I, P, and B frames, and motion prediction is achieved by comparing certain frames with other frames. The I frame is an independent frame, meaning that it can be reconstructed without reference to any other frames. A P (for previous) frame is compared with the previous I frame, and only those parts which differ as a result of movement need to be encoded. The comparison is carried out in sections called *macroblocks* for the frames. A macroblock consists of 16×16 pixels. A B (for bidirectional) frame is compared with the previous I or P frame and with the next P frame. This obviously means that frames must be stored in order for the forward comparison to take place. Only the changes resulting from motion are encoded, which provides further compression. An estimate of the compression required can be made by assuming a value of 200 Mb/s for the uncompressed bit rate for SDTV (see Table 16.1), and taking 5 Mb/s as typical of that for a TV channel, the compression needed is on the order 200/5 = 40:1. The 5 Mb/s would include audio and data, but these should not take more than about 200 kb/s. Audio compression is discussed later in this section.

The whole encoding process relies on digital decision-making circuitry and is computationally intensive and expensive. The decoding process is much simpler because the rules for decoding are part of the syntax of the bit stream. Decoding is carried out in the *integrated receiver decoder* (IRD) unit. This is described in Sec. 16.8.

In DBS systems, MPEG-1 is used for audio compression, and as discussed earlier, MPEG-2 is used for video compression. Both of these MPEG standards cover audio and video, but MPEG-1 video is not designed for DBS transmissions. MPEG-1 audio supports mono and two-channel stereo only, which is considered adequate for DBS systems currently in use. MPEG-2 audio supports multichannel audio in addition to mono and stereo. It is fully compatible with MPEG-1 audio, so the IRDs, which carry MPEG-2 decoders, will have little trouble in being upgraded to work with DBS systems transmitting multichannel audio.

The need for audio compression can be seen by considering the bit rate required for high-quality audio. The bit rate is equal to the number of samples per second (the sampling frequency f_s) multiplied by the number of bits per sample n:

$$R_b = f_s \times n \tag{16.3}$$

For a stereo CD recording, the sampling frequency is 44.1 kHz, and the number of bits per sample is 16:

$$R_b = 44.1 \times 10^3 \times 16 \times 2 = 1411.2 \text{ kb/s}$$

The factor 2 appears on the right-hand side because of the two channels in stereo. This bit rate, approximately 1.4 Mb/s, represents too high a fraction of the total bit rate allowance per channel, and hence the need for audio compression. Audio compression in MPEG exploits certain *perceptual phenomena* in the human auditory system. In particular, it is known that a loud sound at one particular frequency will mask a less intense sound at a nearby frequency. For example, consider a test conducted using two tones, one at 1000 Hz, which will be called the *masking tone*, and the other at 1100 Hz, the *test tone*. Starting with both tones at the same level, say, 60 dB above the threshold of hearing, if now the level of the 1000-Hz tone is held constant while reducing the level of the 1100-Hz tone, a point will be reached where the 1100-Hz tone becomes inaudible. The 1100-Hz tone is said to be *masked* by the 1000-Hz tone. Assume for purposes of illustration that the test tone becomes inaudible when it is 18 dB below the level of the masking tone. This 18 dB is the *masking threshold*. It follows that any noise below the masking threshold also will be masked.

For the moment, assuming that only these two tones are present, then it can be said that the *noise floor* is 18 dB below the masking tone. If the test-tone level is set at, say, 6 dB below the masking tone, then of course it is 12 dB above the noise floor. This means that the signal-to-noise ratio for the test tone need be no better than 12 dB. Now in a *pulse-code modulated* (PCM) system the main source of noise is that arising from the quantization process (see Sec. 10.3). It can be shown (see Roddy and Coolen, 1995) that the signal-to-quantization noise ratio is given by

$$\left(\frac{S}{N}\right)_q = 2^{2n} \tag{16.4}$$

where n is the number of bits per sample. In decibels this is

$$\begin{aligned}\left[\frac{S}{N}\right]_q &= 10 \log 2^{2n} \\ &\cong 6n \text{ dB}\end{aligned} \tag{16.5}$$

This shows that increasing n by 1 bit increases the signal-to-quantization noise ratio by 6 dB. Another way of looking at this is to say that a 1-bit decrease in n increases the quantization noise by 6 dB. In the example above where 12 dB is an adequate signal-to-noise ratio, Eq. (16.5) shows that only 2 bits are needed to encode the 1100-Hz tone (i.e., the levels would be quantized in steps represented by 00, 01, 10, 11). By way of contrast, the CD samples taken at a sampling frequency of 44.1 kHz are quantized using 16 bits to give a signal-to-quantization noise ratio of 96 dB.

Returning to the example of two tones, in reality, the audio signal will not consist of two single tones but will be a complex signal covering a wide spectrum of frequencies. In MPEG-1, two processes take place in parallel, as illustrated in Fig. 16.4. The filter bank divides the spectrum of the incoming signal into subbands. In parallel with this the spectrum is analyzed to permit identification of the masking levels. The masking information is passed to the quantizer, which then quantizes the subbands according to the noise floor.

The masking discussed so far is referred to as *frequency masking* for the reasons given earlier. It is also an observed phenomenon that the masking effect lasts for a short period after the masking signal is removed. This is termed *temporal masking*, and it allows further compression in that it extends the time for which the reduction in quantization applies. There are many more technical details to MPEG-1 than can be covered here, and the reader is referred to Mead (2000), which contains a detailed analysis of MPEG-1. The MPEG Web page—at http://www.mpeg.org/MPEG/audio.html—also provides a number of articles on the subject. The compressed bit rate for MPEG-1 audio used in DBS systems is 192 kb/s.

MPEG-4 (part 10) was developed jointly by the *Video Coding Experts Group* (VCEG) of the *International Telecommunication Union* (ITU), Telecommunication Standardization Sector (ITU-T) which uses the designation H.264, and the MPEG of the ISO/IEC. As noted in Sullivan et al. (2004), this version of MPEG is known by at least six different

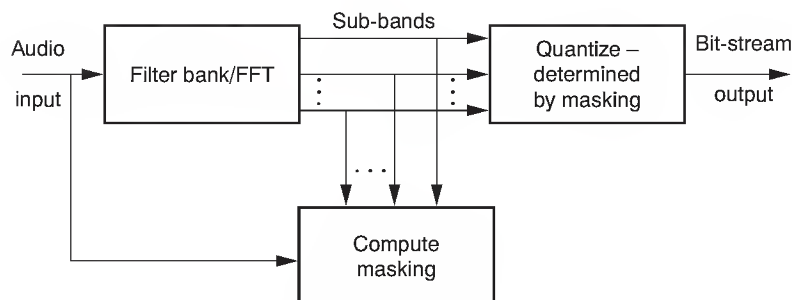


Figure 16.4 MPEG-1 block schematic.

names (H.264, H.26L, ISO/IEC 14496-10, JVT, MPEG-4 AVC, and MPEG-4 Part 10) and the abbreviation AVC is commonly used to denote *advanced video coding*. Following the usage in Sullivan et al., it will be denoted here by H.264/AVC.

Areas of application include video telephony, video storage and retrieval (DVD and hard disk), digital video broadcast, and others. In general terms, MPEG-4 provides many features not present with other compression schemes, such as interactivity for viewers, where objects within a scene can be manipulated, but from the point of view of satellite television, the major advantage is the reduction in bit rate offered. About a 2:1 reduction in bit rate, on average is achievable with H.264/AVC compared with MPEG-2, and in July 2004, an amendment known as *fidelity range extensions* (FRExt, amendment 1) was added to H.264/AVC that can provide a reduction of as much as 3:1 in certain situations (Sullivan et al., 2004). FRExt supports 4:2:2 and 4:4:4 sampling.

As with MPEG-2 the analog outputs from the red (R), green (G), and blue (B) color cameras are converted to a luminance component (Y) and two chrominance components (Cr) and (Cb) but with a different **M** matrix, this being:

$$\begin{bmatrix} \text{Y} \\ \text{Cr} \\ \text{Cb} \end{bmatrix} = \begin{bmatrix} 0.2126 & 0.587 & 0.0722 \\ -0.119977 & -0.331264 & 0.523589 \\ 0.561626 & -0.418688 & -0.051498 \end{bmatrix} \begin{bmatrix} \text{R} \\ \text{G} \\ \text{B} \end{bmatrix} \tag{16.6}$$

It follows that any format conversion would require a matrix recalculation.

H.264/AVC takes advantage of the increases in processing power available from computer chips, but at the cost of more expensive equipment, both for the TV broadcaster and the consumer. As with MPEG-2, frames are compared for changes through comparing macroblocks of 16×16 pixels, but H.264/AVC also allows for comparisons of submacroblocks of pixel groups 16×8, 8×16, 8×8, 8×4, 4×8, and 4×4. At present it is not backward compatible with MPEG-2, which may present a problem with some high definition TV (see Sec. 16.13).

16.8 Forward Error Correction (FEC)

Because of the highly compressed nature of the DBS signal, there is little redundancy in the information being transmitted, and bit errors affect the signal much more severely than they would in a noncompressed bit stream. As a result, FEC is a must. Concatenated coding is used (see Sec. 11.6). The outer code is a Reed-Solomon code that corrects for block errors, and the inner code is a convolution code that corrects for random errors. The inner decoder utilizes the Viterbi decoding algorithm. These FEC bits, of course, add overhead to the bit stream. Typically, for a

240-W transponder (see Sec. 16.3), the bit capacity of 40 Mb/s (see Sec. 16.5) may have a payload of 30 Mb/s and coding overheads of 10 Mb/s. The lower-power 120-W transponders require higher coding overheads to make up for the reduction in power and have a payload of 23 Mb/s and coding overheads of 17 Mb/s. More exact figures are given in Mead (2000) for DirecTV, where the overall code rates are given as 0.5896 for the 120-W transponder and 0.758 for the 240-W transponder.

Mead (2000) has shown that with FEC there is a very rapid transition in BER for values of $[E_b/N_0]$ between 5.5 and 6 dB. For $[E_b/N_0]$ values greater than 6 dB, the BER is negligible, and for values less than 5.5 dB, the BER is high enough to render the system useless.

The use of turbo codes, and LDPC codes (Sec. 11.11) currently being introduced for high definition TV will provide a much greater increase in transponder capacity. As mentioned in Sec. 11.11.1, the Digital Video Broadcast S2 standard (DVB-S2) employs LDPC as the inner code in its FEC arrangement (Breynaert, 2005, Yoshida, 2003), and the DVB-RCS plans to use turbo codes (talk Satellite, 2004).

16.9 The Home Receiver Outdoor Unit (ODU)

The home receiver consists of two units—an outdoor unit and an indoor unit. Commercial brochures refer to the complete receiver as an IRD. Figure 16.5 is a block schematic for the ODU. This will be seen to be similar to the outdoor unit shown in Fig. 8.1. The downlink signal, covering the frequency range 12.2 to 12.7 GHz, is focused by the antenna into the receive horn. The horn feeds into a polarizer that can be switched to pass either left-hand circular or right-hand circular polarized signals. The low-noise block that follows the polarizer contains a *low-noise amplifier* (LNA) and a downconverter. The function of the LNA is described in Sec. 12.5. The downconverter converts the 12.2- to 12.7-GHz band to 950 to 1450 MHz, a frequency range better suited to transmission through the connecting cable to the indoor unit.

The antenna usually works with an offset feed (see Sec. 6.14), and a typical antenna structure is shown in Fig. 16.6. It is important that the antenna have an unobstructed view of the satellite cluster to which it is aligned. Alignment procedures are described in Sec. 3.2.

The size of the antenna is a compromise among many factors but typically is around 18 in. (46 cm) in diameter. A small antenna is desirable for a number of reasons. Small antennas are less intrusive visually and also are less subject to wind loading. In manufacture, it is easier to control surface irregularities, which can cause a reduction in gain by scattering the signal energy. The reduction can be expressed as a function of the *root-mean-square* (rms) deviation of the surface, referred to an *ideal parabolic surface*.

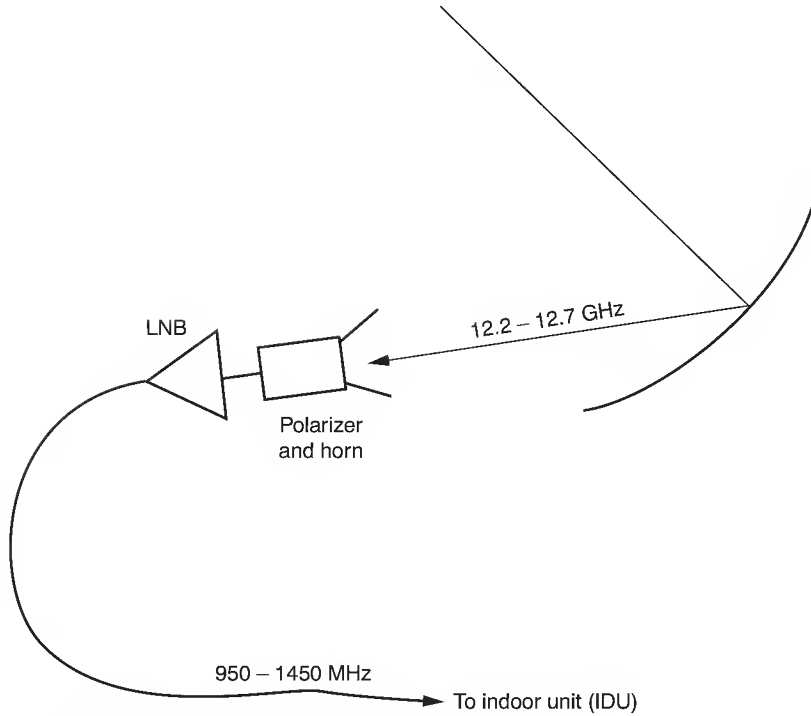


Figure 16.5 Block schematic for the outdoor unit (ODU).

The reduction in gain is given by (see Baylin and Gale, 1991)

$$\eta_{\text{rms}} = e^{8.8\sigma/\lambda} \tag{16.7}$$

where σ is the rms tolerance in the same units as λ, the wavelength. For example, at 12 GHz (wavelength 2.5 cm) and for an rms tolerance of 1 mm, the gain is reduced by a factor

$$\eta_{\text{rms}} = e^{8.8\times 0.1/2.5} = 0.7$$

This is a reduction of about 1.5 dB.

The isotropic power gain of the antenna is proportional to $(D/\lambda)^2$, as shown by Eq. (6.32), where D is the diameter of the antenna. Hence, increasing the diameter will increase the gain (less any reduction resulting from rms tolerance), and in fact, many equipment manufacturers provide signal-strength contours showing the size of antenna that is best for a given region. Apart from the limitations to size stated earlier, it should be noted that at any given DBS location there are *clusters* of satellites, as described in Sec. 16.2. The beamwidth of the antenna must be wide enough to receive from all satellites in the cluster. For example,



Figure 16.6 A typical DBS antenna installation.

the Hughes DBS-1 satellite, launched on December 18, 1993, is located at 101.2°W longitude; DBS-2, launched on August 3, 1994, is at 100.8°W longitude; and DBS-3, launched on June 3, 1994, is at 100.8°W longitude. There is a spread of plus or minus 0.2° about the nominal 101°W position. The −3-dB beamwidth is given as approximately $70\lambda/D$, as shown by Eq. (6.33). A 60-cm dish at 12 GHz would have a −3-dB beamwidth of approximately $70 \times 2.5/60 = 2.9°$, which is adequate for the cluster.

16.10 The Home Receiver Indoor Unit (IDU)

The block schematic for the IDU is shown in Fig. 16.7. The transponder frequency bands shown in Fig. 16.2 are downconverted to be in the range 950 to 1450 MHz, but of course, each transponder retains

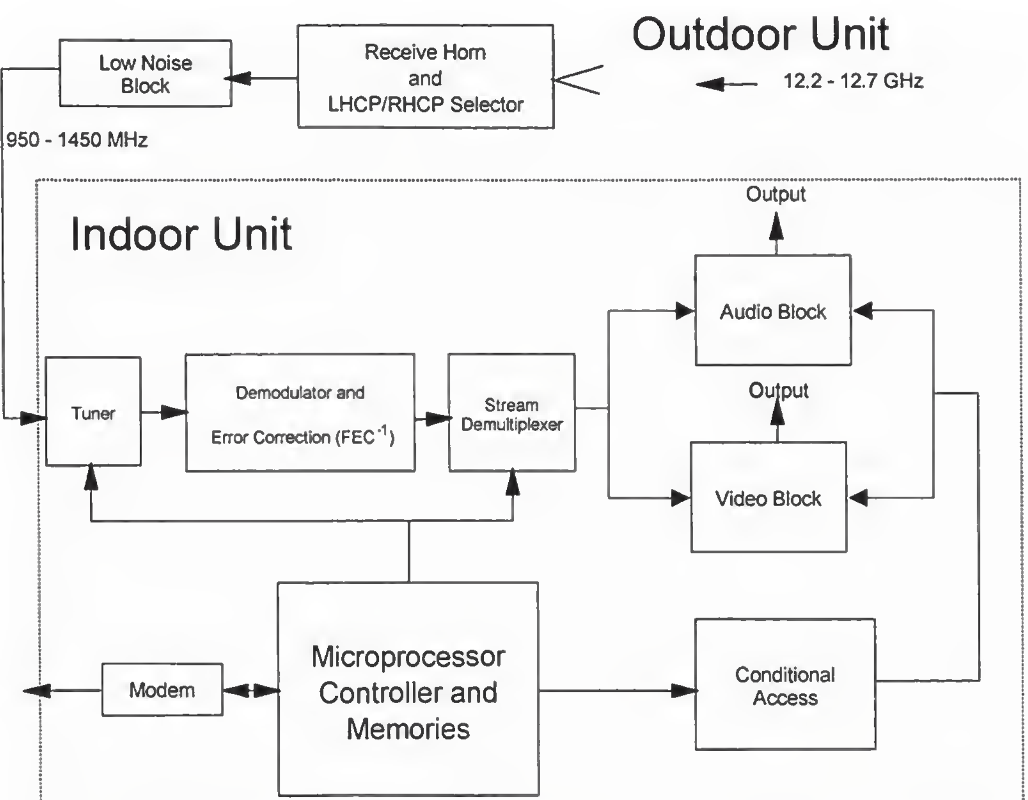


Figure 16.7 Block schematic for the indoor unit (IDU).

its 24-MHz bandwidth. The IDU must be able to receive any of the 32 transponders, although only 16 of these will be available for a single polarization. The tuner selects the desired transponder. It should be recalled that the carrier at the center frequency of the transponder is QPSK modulated by the bit stream, which itself may consist of four to eight TV programs TDM. Following the tuner, the carrier is demodulated, the QPSK modulation being converted to a bit stream. Error correction is carried out in the decoder block labeled FEC^{-1}. The demultiplexer following the FEC^{-1} block separates the individual programs, which are then stored in buffer memories for further processing (not shown in the diagram). This further processing would include such things as conditional access, viewing history of *pay-per-view* (PPV) usage, and connection through a modem to the service provider (for PPV billing purposes). A detailed description of the IRD will be found in Mead (2000).

16.11 Downlink Analysis

The main factor governing performance of a DBS system will be the $[E_b/N_0]$ of the downlink. The downlink performance for clear-sky conditions can be determined using the method described in Sec. 12.8 and illustrated in Example 16.1 that follows. The effects of rain can be calculated using the procedure given in Sec. 12.9.2 and illustrated in Example 16.2 that follows. In calculating the effects of rain, use is made of Fig. 16.8, which shows the regions (indicated by letters) tabulated along with rainfall in Table 16.2.

Example 16.1 A ground station located at 45°N and 90°W is receiving the transmission from a DBS at 101°W. The [EIRP] is 55 dBW, and the downlink frequency may be taken as 12.5 GHz for calculations. Transmission at the full capacity of 40 Mbps may be assumed. An 18-in-diameter antenna is used, and an efficiency of 0.55 may be assumed. Miscellaneous transmission losses of 2 dB also may be assumed. For the IRD, the equivalent noise temperature at the input to the LNA is 100 K, and the antenna noise temperature is 70 K. Calculate the look angles for the antenna, the range, and the $[E_b/N_0]$ at the IRD and comment on this.

Solution Use a value of 6371 km for the mean earth radius, and 42164 km for the geostationary radius.

From Eq. (3.8):

$$B = \phi_E - \phi_{SS} = -90° - (-101°) = 11°$$

From Eq. (3.9):

$$b = a\cos(\cos B \cdot \cos \lambda_E) = a\cos(\cos 11° \cdot \cos 45°) = 46.04°$$

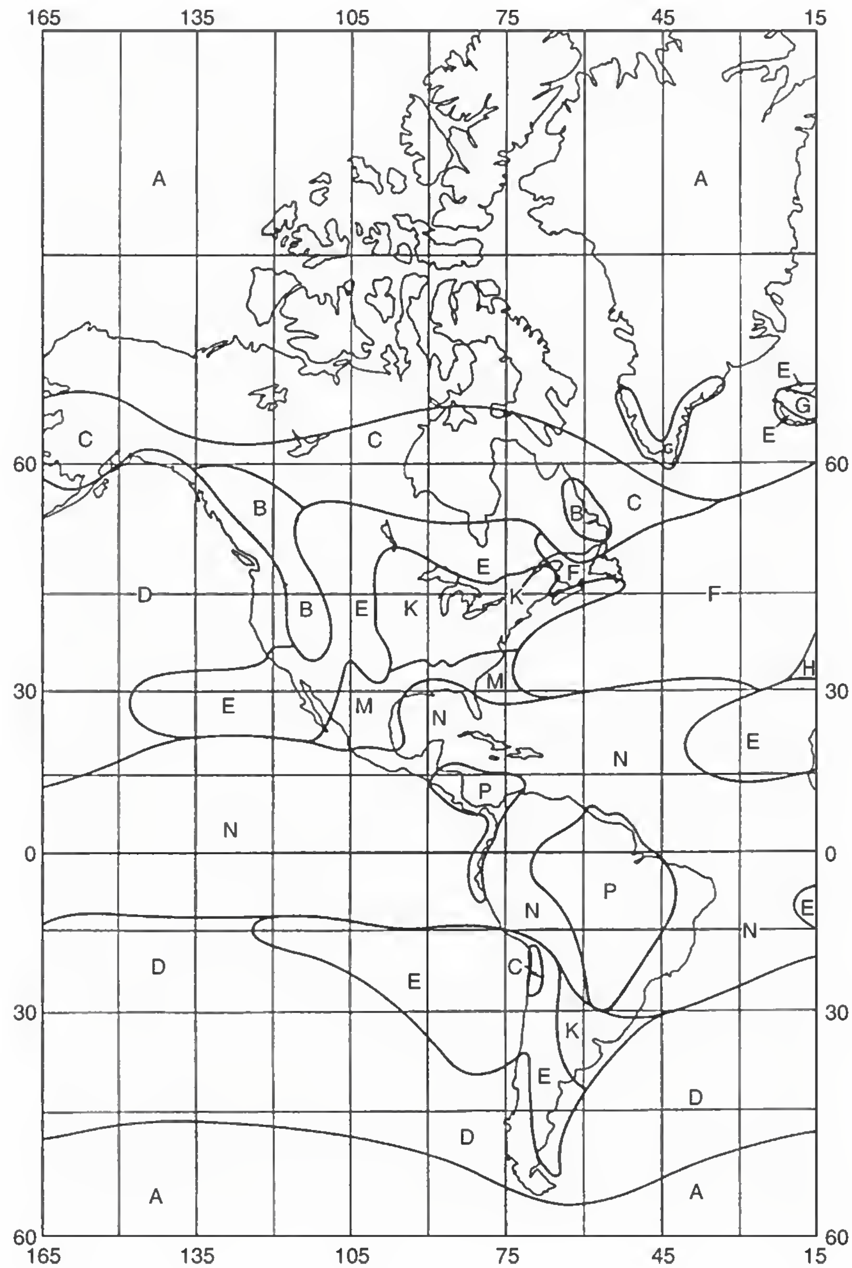


Figure 16.8 Rain climatic zones. Refer to Table 16.2. (*Courtesy of Rec. ITU-R PN.837-1, with permission from the copyright holder ITU. Sole responsibility for the reproduction rests with the author. The complete volume of the ITU material from which the material is extracted can be obtained from the International Telecommunication Union, Sales and Marketing Service, Place des Nations–CH-1211, Geneva 20, Switzerland.*)

TABLE 16.2 Rainfall Intensity Exceeded (mm/h) (Refer to Fig. 16.8)

Percentage of time	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q
1.0	<0.1	0.5	0.7	2.1	0.6	1.7	3	2	8	1.5	2	4	5	12	24
0.3	0.8	2	2.8	4.5	2.4	4.5	7	4	13	4.2	7	11	15	34	49
0.1	2	3	5	8	6	8	12	10	20	12	15	22	35	65	72
0.03	5	6	9	13	12	15	20	18	28	23	33	40	65	105	96
0.01	8	12	15	19	22	28	30	32	35	42	60	63	95	145	115
0.003	14	21	26	29	41	54	45	55	45	70	105	95	140	200	142
0.001	22	32	42	42	70	78	65	83	55	100	150	120	180	250	170

SOURCE: Table 16.2 and Fig. 16.8 reproduced from ITU Recommendation ITU-R PN.837-1 (1994), with permission.

From Eq. (3.10):

$$A = a\sin\left(\frac{\sin|B|}{\sin b}\right) = 15.37°$$

By inspection, $\lambda_E > 0$; therefore, Fig. 3.3*d* applies and the required azimuth angle is:

$$A_Z = 180° + A = \underline{\underline{195.37°}}$$

From Eq. (3.11):

$$\begin{aligned} d &= \sqrt{R^2 + a_{\text{GSO}}^2 - 2 \cdot R \cdot a_{\text{GSO}} \cos b} \\ &= \sqrt{6371^2 + 42164^2 - 2 \times 6371 \times 42164 \times \cos 46.04} \\ &= 38019.1 \text{ km} \end{aligned}$$

Equation (3.12) gives the required angle of elevation as

$$El = a\cos\left(\frac{a_{\text{GSO}}}{d}\sin b\right) = a\cos\left(\frac{42164}{38019.1}\sin 46.04°\right) \cong \underline{\underline{37°}}$$

As noted in Sec. 3.2, the calculated values for azimuth and elevation provide a guide. Practical adjustments would be made to maximize the received signal.

Equation (12.10), with d in km and f in MHz gives:

$$[\text{FSL}] = 32.4 + 20\log d + 20\log f = 205.94 \text{ dB}$$

Adding in the miscellaneous losses, Eq. (12.12) gives the total losses as

$$[\text{LOSSES}] = [\text{FSL}] + 2 = 207.94 \text{ dB}$$

The system noise temperature, from Eq. (12.23) is

$$T_S = T_{\text{eq}} + T_{\text{ant}} = 100 + 70 = 170 \text{ K}$$

In decibels relative to 1° K this is

$$[T_S] = 10\log T_S = 22.3 \text{ dBK}$$

From Eq. (12.5), with f in GHz and D in ft:

$$G = \eta(3.192 \times f \times D)^2 = 0.55 \times (3.192 \times 12.5 \times 1.5)^2 = 1970.1$$

In decibels this is

$$[G] = 10\log 1970.1 = 32.94 \text{ dB}$$

From Eq. (12.35):

$$\left[\frac{G}{T}\right] = [G] - [T_S] = 10.64 \text{ dBK}^{-1}$$

Equation (12.53) gives

$$\left[\frac{C}{N_0}\right] = [\text{EIRP}] + \left[\frac{G}{T}\right] - [\text{LOSSES}] - [K]$$

$$= 55 + 10.64 - 207.94 + 228.6$$

$$= 86.3 \text{ dBHz}$$

The downlink bit rate in decibels relative to 1 bps is

$$[R_b] = 10\log(40 \times 10^6) \cong 76 \text{ dBbps}$$

Equation (10.24) gives:

$$\left[\frac{E_b}{N_0}\right] = \left[\frac{C}{N_0}\right] - [R_b] \cong \underline{\underline{10.3 \text{ dB}}}$$

As noted in Sec. 16.8, a $[E_b/N_0]$ of at least 6 dB is required. The value obtained provides a margin of 4.3 dB under clear-sky conditions.

Example 16.2 Table 16.2 and Fig. 16.8 show the rainfall intensity in mm/h exceeded for given percentages of time. Calculate the upper limit for $[E_b/N_0]$ set by the rainfall for the percentage of time equal to 0.01 percent. The earth station is at mean sea level, and the rain attenuation may be assumed entirely absorptive, and the apparent absorber temperature may be taken as 272 K.

Solution It is first necessary to calculate the attenuation resulting from the rain. The given data are shown below. Because the CCIR formula contains hidden conversion factors, units will not be attached to the data, and it is understood that all lengths and heights are in km, and rain rate is in mm/h. From Fig. 16.8, the earth station is seen to be located within region K. From the accompanying Table 16.2, the rainfall exceeds 42 mm/h for 0.01 percent of the time. Table 4.2 does not give the coefficients for 12.5 GHz; therefore, the values must be found by linear interpolation between 12 and 15 GHz. Denoting the 12-GHz values with subscript 12 and the 15-GHz values with subscript 15, then and using the values from Table 4.2,

$$a_h = a_{h12} + \frac{a_{h15} - a_{h12}}{15 - 12} \times (12.5 - 12)$$

$$= 0.0188 + \frac{0.0367 - 0.0188}{3} \times 0.5$$

$$= 0.0218$$

$$b_h = b_{h12} + \frac{b_{h15} - b_{h12}}{15 - 12} \times (12.5 - 12)$$

$$= 1.217 + \frac{1.154 - 1.217}{3} \times 0.5$$

$$= 1.207$$

$$a_v = a_{v12} + \frac{a_{v15} - a_{v12}}{15 - 12} \times (12.5 - 12)$$

$$= 0.0168 + \frac{0.0335 - 0.0168}{3} \times 0.5$$

$$= 0.0196$$

$$b_v = b_{v12} + \frac{b_{v15} - b_{v12}}{15 - 12} \times (12.5 - 12)$$

$$= 1.2 + \frac{1.128 - 1.2}{3} \times 0.5$$

$$= 1.188$$

Since circular polarization is used, the coefficients are found from Eq. (4.8):

From Eq. (4.8*a*):

$$a_c = \frac{a_h + a_v}{2} = \frac{0.0218 + 0.0196}{2} = 0.0207$$

From Eq. (4.8*b*):

$$b_c = \frac{a_h \cdot b_h + a_v \cdot b_v}{2 \cdot a_c} = \frac{0.0218 \times 1.207 + 0.0196 \times 1.188}{2 \times 0.0207} = 1.198$$

Using the Method 3 curves in Fig. 4.4 for p = 0.01 percent and earth-station latitude 45°, the rain height is approximately 3.5 km, and as stated in the problem, at mean sea level $h_o = 0$

From Eq. (4.4):

$$L_S = \frac{h_R - h_0}{\sin El} = \frac{3.5}{\sin 37°} = 5.82 \text{ km}$$

From Eq. (4.6)

$$L_G = L_S \cos El = 5.82 \times \cos 37° = 4.64 \text{ km}$$

From Table 4.3:

$$r_{01} = \frac{90}{90 + 4L_G} = 0.829$$

From Eq. (4.5):

$$L = L_S \cdot r_{01} = 5.82 \times .829 = 4.82$$

From Eq. (4.2):

$$\alpha = a_c R_{01}^{b_c} = 0.0207 \times 42^{1.198} = 1.819$$

From Eq. (4.3):

$$[A_{01}] = \alpha L = 1.819 \times 4.82 = \underline{\underline{8.76 \text{ dB}}}$$

The effect of rain is calculated as shown in Sec. 12.9.2. This requires the attenuation to be expressed as a power ratio

$$A = 10^{[A_{01}]/10} = 10^{0.876} = 7.52$$

For Eq. (12.60), the noise-to-signal ratios are required. From Example 16.1, $[C/N_0] =$ 86.3 dBHz for clear sky, hence,

$$\left(\frac{C}{N_0}\right)_{\text{CS}} = 10^{-86.3/10} = 2.34 \times 10^{-9}$$

The system noise temperature under clear-sky conditions is just T_S, but the subscript will be changed to conform with Eq. (12.60): $T_{\text{SCS}} = T_S = 170$ K

From Eq. (12.60):

$$\left(\frac{N_0}{C}\right)_{\text{rain}} = \left(\frac{N_0}{C}\right)_{\text{CS}} \left(A + (A - 1) \cdot \frac{T_a}{T_{\text{SCS}}}\right)$$

$$= 2.34 \times 10^{-9}\left(7.52 + 6.52 \times \frac{272}{170}\right)$$

$$= 4.21 \times 10^{-8}$$

Hence,

$$\left[\frac{C}{N_0}\right]_{\text{rain}} = 10 \log (4.21 \times 10^{-8})^{-1} = 73.76 \text{ dBHz}$$

Recalculating the $[E_b/N_0]$ ratio, Eq. (10.24) gives:

$$\left[\frac{E_b}{N_0}\right]_{\text{rain}} = \left[\frac{C}{N_0}\right]_{\text{rain}} - [R_b] = \underline{\underline{-2.26 \text{ dB}}}$$

Thus the rain will completely wipe out the signal for 0.01 percent of the time. It is left as an exercise for the reader to find the size of antenna that would provide an adequate signal under these rain conditions.

16.12 Uplink

Ground stations that provide the uplink signals to the satellites in a DBS system are highly complex systems in themselves, utilizing a wide range of receiving, recording, encoding, and transmission equipment. Signals will originate from many sources. Some will be analog TV received from satellite broadcasts. Others will originate in a studio, others from video cassette recordings, and some will be brought in on cable or optical fiber. Data signals and audio broadcast material also may be included. All of these must be converted to a uniform digital format, compressed, and time division multiplexed. Necessary service additions which must be part of the multiplexed stream are the program guide and conditional access. FEC is added to the bit stream, which is then used to QPSK modulate the carrier for a given transponder. The whole process, of course, is duplicated for each transponder carrier.

Because of the complexity, the uplink facilities are concentrated at single locations specific to each broadcast company. The uplink facilities for Echostar's DISH network are shown in Fig. 16.9. The four uplink



Figure 16.9 Uplink facilities for Echostar's DISH network. (*Courtesy of Echostar, at* http://www.dishnetwork.com/content/aboutus/presskit/print_satellites/index.shtml)

TABLE 16.3 Uplink Facilities in Operation as of 1996

Company	Location
AlphaStar	Oxford, Connecticut
DirecTV	Castle Rock, Colorado
EchoStar	Cheyenne, Wyoming
U.S. Satellite Broadcasting	Oakdale, Minnesota

facilities in operation as of 1996, as given in Mead (2000), are shown in Table 16.3.

16.13 High Definition Television (HDTV)

Table 16.1 shows the 18 *advanced television systems committee* (ATSC) formats for digital television, which includes HDTV. Some of the early HDTV systems were analog, and a description of these will be found in Kuhn, (1995). However, all analog TV transmissions in the U.S. are scheduled to shut down by January 1, 2009. It will be possible to get "set-top boxes" that convert HDTV signals to a format suitable for analog sets, but the high resolution, and many of the other advantages of digital TV will be lost in this process. In Europe, Astra is the major satellite provider, and they, along with 60 European broadcasters have agreed to standardize on two HDTV formats: 720p50 meaning 720 lines (or pixels) of vertical resolution, with progressive scan, at a 50 Hz refresh rate and 1080i25, meaning 1080 lines (or pixels) of vertical resolution, with interlace scan, at a 25 Hz refresh rate. The refresh rates are normally tied in with the frequency of the domestic electricity supply, 60 Hz in the U.S. and most of north America, and 50 Hz in Europe. In Table 16.1 a refresh rate of 24 Hz is listed, to make this format compatible with film projection at 24 frames per second.

DirecTV plans to use H.264/AVC (MPEG-4 Part 10, see Sec. 16.7) in its HDTV satellite broadcasts and all HDTV services in Europe are expected to use this rather than the MPEG-2. Comparing H.264/AVC to MPEG-2 the bit rate reduction can vary between about 50 (Mark, 2005) to 65 percent (Wipro, 2004). Two high definition channels require a bit rate of 16 to 18 Mbps with MPEG-2 and 6 to 8 Mbps with H.264/AVC. As explained in Sec. 16.7, H.264/AVC is not backward compatible with MPEG-2, and this may mean considerable expense for the consumer who wishes to receive HDTV along with SDTV. The MPEG is working on making H.264/AVC compatible with MPEG-2 (Koenen, 1999).

16.13.1 HDTV displays

The familiar direct view *cathode ray tube* (CRT) used for analog TV is not capable of displaying HDTV. Rear projection CRT (RPTV) sets are probably the least expensive of the "big screen" sets suitable for HDTV,

although these are being replaced by newer technology. Among the competing technologies are plasma displays, *liquid crystal displays* (LCD) and *digital light processing* (DLP) displays (there are others, but these are the most prominent ones). Plasma displays are made up of tiny cells coated with red, green, and blue phosphors. The video signal stimulates a gas inside the cells, which impacts the phosphors causing them to glow. Plasma flat-panel displays are around 3 to 5 in. thick and screen sizes up to 60 in. are available. In a LCD light passes through a thin sheet of the liquid crystal material which forms the viewing screen. A thin film transistor array carrying the video signal produces varying degrees of polarization of the liquid crystal allowing more or less light to pass through for each of the colors red, green, and blue. LCDs are thin, flat panel displays that can be made with screen sizes up to about 50 in. DLP displays utilize what is known as a *digital micromirror device* (DMD) invented by Texas Instruments. The DMD contains approximately 1.3 million micro-mirrors, each micro-mirror representing one pixel. The micro-mirrors can be mechanically pivoted up to 5000 times a second, the pivoting being activated by the video signal. The degree of pivoting is determined by the signal level, and light reflected by the DMD is passed through a "color wheel" consisting of red, blue, and green filters, which rotates at speeds of about 120 revolutions per second. Each filter projects an image for a brief period onto the screen and the eye integrates these to "see" the composite picture. The DLP display is a rear projection unit and is about 12 to 14 in. deep. These units are available in large screen size. LCD and DLP displays require a light source; they can also be made as front projection units, which provide corresponding larger screens than rear projection displays. Apart from the CRT, all of these displays can be activated by digital signals, thus avoiding the need for digital to analog conversion.

Although the vision aspects of HDTV are by nature the most noticeable, sound quality is also very high. HDTV provides for Dolby 5.1 sound, which means there are 5 channels: left; right; center; left rear; right rear; the .1 stands for a sub-woofer, a very low frequency, or bass channel. The audio signal format is Dolby Digital/AC-3.

16.14 Video Frequency Bandwidth

An estimate of the highest frequency in the analog video signal can be found as follows. Let L_{act} be the number of active lines per frame, and L_{supp} the number of lines suppressed during picture flyback, then the total number of lines per frame is $L = L_{act} + L_{supp}$. Each pixel is the width of a line, so the number of lines also represents the number of pixels along the picture viewing height. Let h be the (viewing) picture height. The height of a pixel is therefore $h_{pix} = h/L_{act}$. Let w be the (viewing) picture width, then the number of active pixels per line is

$PL_{act} = w/h_{pix} = (w/h)L_{act}$. But part of the total line is suppressed to allow for line flyback. Let the ratio of total line scan time to active scan time be F_{lfb}, then the total number of pixels per line is $PL_{tot} = PL_{act}F_{lfp} = (w/h)L_{act}F_{lfp}$. The total number of pixels per frame is, therefore, $PF_{tot} = L(w/h)L_{act}F_{lfp}$. An estimate of the highest frequency can be made by assuming that it occurs when a frame consists of alternate black and white pixels such that two pixels make up one cycle. Thus multiplying half the total number of pixels by the frame rate (in frames per second) gives the highest frequency. Subjective tests have shown that a reduction in the highest frequency obtained in this way can be tolerated, the reduction being accounted for by the *Kell factor K*. K = 0.7 is often assumed in practice, although different values are sometimes encountered. In a table of values published by Evans Associates (http://www.evansassoc.com) the Kell factor is given as 0.7 for interlaced scanning and 0.9 for progressive scanning.

The aspect ratio is defined as $a = w/h$ and denoting the number of frames per second by F, the highest frequency is given by

$$\begin{aligned} f_{max} &= \frac{KPF_{tot}F}{2} \\ &= \frac{K \cdot a \cdot L \cdot L_{act} \cdot F_{lfb} \cdot F}{2} \end{aligned} \tag{16.8}$$

As shown by Eq. (16.8) the highest frequency is directly proportional to frame rate. The frame rate is tied to half the frequency of the domestic electricity supply, 60 Hz in the United States and most of North America, and 50 Hz in Europe. At 30 complete frames per second (and more noticeably at 25 frames per second in European systems) flicker was apparent in analog TV. To overcome this, without increasing the frame rate (and hence the highest frequency) frames were divided into two fields, one field consisting of odd numbered lines, the other of even numbered lines. This is termed *interlaced* scanning. Displaying the odd and even fields alternately at 60 fields per second (50 in Europe) keeps the number of frames at 30 per second (25 in Europe) which eliminates the flicker without any increase in the video frequency. With *progressive* scanning the lines are scanned in sequence which provides a sharper picture.

Example 16.3 For NTSC analog TV, (see Sec. 9.5) $L_{act} = 483$, $L = 525$, a 4/3 $F_{lfb} = 1.19$ and $F = 30$. With $K = 0.7$. Calculate the highest video frequency.

Solution Substituting the given values in Eq. (16.8) gives:

$$\begin{aligned} f_{max} &= \frac{.7}{2} \times \frac{4}{3} \times 525 \times 483 \times 1.19 \times 30 \\ &\cong \underline{\underline{4.2 \text{ MHz}}} \end{aligned}$$

Values applicable to digital TV are shown in Table 16.4.

TABLE 16.4 Digital TV Parameters as used in Eq. (16.8)

No.	Format type	L	L_{act}	F_{lfb}	Aspect ratio a	F, Frames per second
1	SDTV	525	480i	1.13	4:3	30
2	EDTV	525	480p	1.13	4:3	24
3	EDTV	525	480p	1.13	4:3	30
4	EDTV	525	480p	1.13	4:3	60
5	EDTV	525	480i	1.13	4:3	30
6	EDTV	525	480p	1.13	4:3	24
7	EDTV	525	480p	1.13	4:3	30
8	EDTV	525	480p	1.13	4:3	60
9	EDTV	525	480i	1.22	16:9	30
10	EDTV	525	480p	1.22	16:9	24
11	EDTV	525	480p	1.22	16:9	30
12	EDTV	525	480p	1.22	16:9	60
13	HDTV	750	720p	1.29	16:9	24
14	HDTV	750	720p	1.29	16:9	30
15	HDTV	750	720p	1.29	16:9	60
16	HDTV	1125	1080i	1.15	16:9	30
17	HDTV	1125	1080p	1.15	16:9	24
18	HDTV	1125	1080p	1.15	16:9	30

NOTES: HDTV—high-definition television; SDTV—standard definition television; EDTV—enhanced definition television; p—progressive scanning; i—interlaced scanning.
SOURCES: Booth, 1999; www.timefordvd.com, 2004; Maxim/Dallas, 2001.

Example 16.4 Calculate the highest video frequency for HDTV 1080*i*. Assume a Kell factor of 0.7.

$$f_{max} = \frac{0.7}{2} \times \frac{16}{9} \times 1125 \times 1080 \times 1.15 \times 30$$

$$\cong \underline{\underline{26 \text{ MHz}}}$$

16.15 Problems and Exercises

16.1. Referring to Fig. 16.1, calculate (*a*) the total number of transponders broadcasting from each of the orbital positions shown and (*b*) the total number of transponders in use by each service provider.

16.2. The [EIRP] of a 240-W transponder is 57 dBW. Calculate the approximate gain of the antenna.

16.3. The transponder in Prob. 16.2 is switched to 120 W. Calculate the new [EIRP], assuming that the same antenna is used.

16.4. Draw accurately to scale the transponder frequency plan for the DBS transponders 5, 6, and 7 for uplink and downlink.

16.5. Calculate the total bit rate capacity available at each of the orbital slots for each of the service providers listed in Fig. 16.1. State any assumptions made.

16.6. Calculate the bandwidth required to transmit a SDTV format having a resolution of 704×480 pixels at 30 frames per second.

16.7. The R, G, and B colors in Eq. (16.2) are restored by the matrix multiplication

$$\begin{bmatrix} \mathrm{R} \\ \mathrm{G} \\ \mathrm{B} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1.402 \\ 1 & -0.344136 & -0.714136 \\ 1 & 1.772 & 0 \end{bmatrix} \begin{bmatrix} \mathrm{Y} \\ \mathrm{Cr} \\ \mathrm{Cb} \end{bmatrix}$$

where the 3×3 matrix is $\mathbf{M}^{-1}$ [the inverse of $\mathbf{M}$ in Eq. (16.2)]. Verify that this matrix multiplication is correct.

16.8. Briefly describe the video compression process used in MPEG-2.

16.9. Explain what is meant by *masking* in the context of audio compression. Describe how MPEG-1 utilizes the phenomenon of masking to achieve compression.

16.10. Using the overall codes rates (Mead, 2000) given in Sec. 16.8 and assuming a 40-Mb/s transponder, calculate the payload bit rate and the overhead for the low-power and high-power transponders.

16.11. Plot antenna gain as a function of diameter for a paraboloidal reflector antenna for antenna diameters in the range 45 to 80 cm. Use a frequency of 12.5 GHz.

16.12. Assuming that the rms tolerance can be held to 0.2 percent of the diameter in antenna manufacture, calculate the reduction in gain that can be expected with antennas of diameter (*a*) 46 cm, (*b*) 60 cm, and (c) 80 cm.

16.13. Calculate the gain and -3-dB beamwidth for antennas of diameter 18, 24, and 30 in. at a frequency of 12.5 GHz. Assume that the antenna efficiency is 0.55 and that the rms manufacturing tolerance is 0.25 percent of diameter.

16.14. A DBS home receiver is being installed at a location 40°N, 75°W to receive from a satellite cluster at 61.5°. Calculate the look angles for the antenna. It is hoped to use an 18-in antenna, the antenna efficiency being 0.55, and the effect of surface irregularities may be ignored. The system noise temperature is 200 K. The downlink frequency may be taken as 12.5 GHz, the [EIRP] as 55 dBW, and the transponder bit rate as 40 Mb/s. Miscellaneous transmission losses may be ignored. Calculate the received clear sky $[E_b/N_0]$, and state whether this will make for satisfactory reception.

16.15. Following the procedure given in Example 16.2, calculate the rain attenuation for 0.01 percent of time and the corresponding $[E_b/N_0]$ for the system in Prob. 16.14. The ground station may be assumed to be at mean sea level. State if satisfactory reception occurs under these conditions.

16.16. A DBS home receiver is being installed at a location 60°N, 155°W to receive from a satellite cluster at 157°. Calculate the look angles for the antenna. It is hoped to use an 18-in. antenna, the antenna efficiency being 0.55, and the effect of surface irregularities may be ignored. The system noise temperature is 200 K. The downlink frequency may be taken as 12.5 GHz, the [EIRP] as 55 dBW, and the transponder bit rate as 40 Mb/s. Miscellaneous transmission losses may be ignored. Calculate the received clear sky $[E_b/N_0]$, and state whether this will make for satisfactory reception.

16.17. Following the procedure given in Example 16.2, calculate the rain attenuation for the 0.01 percent of time, and the corresponding $[E_b/N_0]$ for the system in Prob. 16.16. The ground station may be assumed to be at mean sea level. State if satisfactory reception occurs under these conditions.

16.18. A DBS home receiver is being installed at a location 15°S, 50°W to receive from a satellite cluster at 61.5°. Calculate the look angles for the antenna. Calculate also the diameter of antenna needed to provide a 5-dB margin approximately on the received $[E_b/N_0]$ under clear-sky conditions. An antenna efficiency of 0.55 may be assumed, and the effect of surface irregularities may be ignored. The antenna noise temperature is 70 K, the equivalent noise temperature at the input of the LNA is 120 K, and miscellaneous transmission losses are 2 dB. The downlink frequency may be taken as 12.5 GHz, the [EIRP] as 55 dBW, and the transponder bit rate as 40 Mb/s.

16.19. For the system described in Prob. 16.18, determine the rainfall conditions that will just cause loss of signal. The ground station may be assumed to be at mean sea level.

16.20. For HDTV the R, G, and B colors in Eq. (16.6) are restored by the matrix multiplication

$$\begin{bmatrix} \mathrm{R} \\ \mathrm{G} \\ \mathrm{B} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1.402 \\ 1.218399 & -0.217953 & -0.507777 \\ 1 & 1.772 & 0 \end{bmatrix} \begin{bmatrix} \mathrm{Y} \\ \mathrm{Cr} \\ \mathrm{Cb} \end{bmatrix}$$

where the 3×3 matrix is $\mathbf{M}^{-1}$ [the inverse of $\mathbf{M}$ in Eq. (16.6)]. Verify that this matrix multiplication is correct.

16.21. Using the data from Table 16.4 calculate the highest video frequency for formats Nos. 12 and 15. Use a Kell factor of 0.7 for No. 12 and 0.9 for No. 15.

References

Baylin, F., and B. Gale. 1991. *Ku-Band Satellite TV.* Baylin Publications, Boulder, Colorado.

Bhatt, B., David B., and D. Hermreck. 1997. "Digital Television: Making It Work." *IEEE Spectrum*, Vol. 34, No. 10, October, pp. 19–28.

Booth, S. A. 1999. "Digital TV in the U.S." *IEEE Spectrum*, Vol. 36, No. 3, March, pp. 39–46.

Breynaert, D., 2005. Analysis of the bandwidth efficiency of DVB-S2 in a typical data distribution network. CCBN, Beijing, March, at www.newtec.com.sg/news/articles/Beijing2005%20DVBS2%20whitepaper%202feb05%20MDOWaylon.pdf-29

Chaplin, J. 1992. "Development of Satellite TV Distribution and Broadcasting." *Electron. Commun.*, Vol. 4, No. 1, February, pp. 33–41.

Dishnetwork Web page, at http://www.dishnetwork.com/

Koenen, R. 1999. "MPEG-4, Multimedia for Our Time." *IEEE Spectrum*, Vol. 36, No. 2, pp. 26–33. February, at http://www.chiariglione.org/mpeg/tutorials/papers/koenen/mpeg-4.htm

Kuhn, K. J. 1995. HDTV—An Introduction, at http://www.ee.washington.edu/conselec/CE/kuhn/hdtv

Mark, S. 2005. HDTV Pushes Telcos Toward MPEG-4 Light Reading. June 24, at http://www.lightreading.com/document.asp?doc_id=76252

Maxim/Dallas. 2001. Application Note 750, Bandwidth Versus Video Resolution, at http://www.maxim-ic.com/appnotes.cfm/appnote_number/750

Mead, D. C. 2000. *Direct Broadcast Satellite Communications.* Addison-Wesley, Reading, MA.

MPEG Web page, at http://www.mpeg.org/

Netravali, A., and A. Lippman. 1995. "Digital Television: A Perspective." *Proc. IEEE*, Vol. 83, No. 6, June, pp. 834–842.

Prichard, W. L., and M. Ogata. 1990. "Satellite Direct Broadcast." *Proc. IEEE*, Vol. 78, No. 7, July, pp. 1116–1140.

Reinhart, E. E. 1990. "Satellite Broadcasting and Distribution in the United States." *Telcommun. J.*, Vol. 57, No. V1, June, pp. 407–418.

Roddy, D., and J. Coolen. 1995. *Electronic Communications,* 4th ed. Prentice-Hall, Englewood Cliffs, NJ.

Sullivan, G. J., P. Topiwala, and Ajay Luthra, 2004. "The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extensions." *SPIE Conference on Applications of Digital Image Processing XXVII, Special Session on Advances in the New Emerging Standard H.264/AVC*, August. Denver, Colorado.

Sweet, W. 1997. "Chiariglione and the Birth of MPEG." *IEEE Spectrum*, Vol. 34, No. 9, pp. 70–77.

talk Satellite. 2004. Spectra Licensing Group Announces Gilat's Entrance into the Turbo Code Licensing Program for DVB-RCS Satellite Applications. December 15, at www.talksatellite.com/EMEAdoc

Timefordvd.com. 2004. "Digital TV & HDTV Tutorial." (Last update on 3.9.2004).

Wipro. 2004. MPEG-4 AVC/H.264 Video Coding. White paper. September, at http://www.wipro.com/insights/embedded.htm

Yoshida, J. 2003. Hughes goes retro in digital satellite TV coding. EETimesUK, November, at www.eetuk.com/tech/news/OEG20031110S0081

Chapter

17

Satellite Mobile and Specialized Services

17.1 Introduction

The idea that three geostationary satellites could provide communications coverage for the whole of the earth, apart from relatively small regions at the north and south poles, is generally credited to Clarke (1945). The basic idea was sound, but of course the practicalities led to the development of a much more complex undertaking than perhaps was envisioned originally. Technological solutions were found to the many problems that were encountered, and as a result, satellite services expanded into many new areas. Geostationary satellites are still the most numerous and well in excess of three! If an average of 2° spacing is assumed, the geostationary orbit would provide 180 orbital positions, and in practice satellites are often employed in clusters, with more than one satellite at an assigned orbital position. The 180 orbital positions are not occupied in a uniform distribution as in practice satellites are concentrated to serve regions where services are most in demand. In addition to geostationary satellites, much more use is being made of nongeostationary orbits, especially *low earth orbits* (LEOs) and *medium earth orbits* (MEOs). As described in Chap. 15, a service may utilize a combination of geostationary and nongeostationary satellites.

Although the developments in satellites generally have led to a need for larger satellites, a great deal has been happening with small satellites, often referred to as *microsats* (see Sweeting, 1992, and Williamson, 1994). In this chapter, brief descriptions are given of some of these services to illustrate the range of applications now found for satellites.

17.2 Satellite Mobile Services

Although countries in the developed world are well served by global communications, there remain large areas and population groups that have very limited access to telecommunications services. In the United States and some European countries, telephone landline density, measured by the number of phone lines per 100 people, is as much as 30 times higher than in China, India, Pakistan, and the Philippines, and an estimated 3 billion people worldwide have no phone at home (Miller, 1998). Developing a telephone network on the ground, whether wired or cellular, is time-consuming and expensive. Civil infrastructure may need to be installed or upgraded, including roads and utilities such as water and electricity. Once satellites are deployed in orbit, they can provide wide area service for telephone, facsimile, and Internet, on an as-needed basis, without the need for extensive ground facilities.

An Internet search for mobile satellite companies will provide a confusing array of data, with company mergers and filings for bankruptcy frequently being reported. Unfortunately, many of the Internet reports do not have dates and it becomes difficult to determine the current situation, or possible future developments, and only those systems that appear to be well established will be described. Most of the systems that offer telephone services provide the users with dual-mode phones that operate to GSM standards. GSM stands for *global system for mobile communications* (originally Groupe Spe'cial Mobile); it is the most widely used standard for cellular and personal communications. The user frequencies are in the L (1 to 2 GHz) and S (2 to 4 GHz) bands, and where geostationary satellites are employed these require large on-board antennas in the range 100 to 200 m^2.

Asian Cellular System. The Asian Cellular System, or AceS, utilizes one Garuda geosynchronous satellite covering the Asia Pacific area, an area of over 11 million square miles. The footprint ranges from China in the north to Indonesia in the south, and the Philippines and Papua New Guinea in the east to India and Pakistan in the west. The satellite is positioned at 123°E longitude with a variation of plus or minus 3°N and S. It will stabilize at the assigned equatorial longitude variation after 3.7 years in operation. As noted in Sec. 7.5, by placing the satellite in an inclined orbit, the N-S station-keeping maneuvers may be dispensed with. The savings in weight achieved by not having to carry fuel for these maneuvers allows the communications payload to be increased, but this arrangement requires the use of tracking antennas at the ground station network control center.

The population of the regions serviced totals over 3.5 billion The Garuda satellite has capacity for at least 11,000 simultaneous telephone channels, servicing up to 2 million subscribers. The satellites utilize

two 12 meter, L-band antennas (large umbrellalike structures) that generate 140 spot beams. On-board digital switching is provided which routes calls between beams. Subscribers are provided with a dual-mode phone that can be switched between satellite and the GSM modes of operation. Services include voice telephony, Internet connectivity, data, and alerting and paging. The mobile links operate at frequencies in the L-band (uplink 1625.5–1660.5 MHz; downlink 1525–1559 MHz), and the gateway frequencies are in the C-band (uplink 6.425–6.725 GHz; downlink 3.400–3.700 GHz). The gateway terminals provide access to the national telephone networks. A second satellite will be employed to expand the service into western and central Asia, Europe, and northern Africa. Further information can be obtained from the Web site at http://www.acesinternational.com/corporate/index.php

Thuraya. The Thuraya satellite system has geosynchronous satellites located at 44° and 28.5°E longitude, at an inclination of 6.3°, again presumably to dispense with N-S station-keeping maneuvers as described in Sec. 7.5. The first Thuraya satellite was launched on October 21, 2000, and the second on June 10, 2003. A third satellite is being built to increase the system capacity. The Thuraya satellite system serves an area between about 20°W to 100°E longitude and 60°N to 2°S latitude, covering more than 110 countries with a combined population of 2.3 billion. The footprint spans Europe, North, Central Africa, and large parts of Southern Africa. A 12.25 × 16 m elliptical antenna is employed providing 250 to 300 spot beams, with onboard beam-switching. An uplink beacon signal provides beam adjustment to compensate for the nongeostationary path of the satellites. The system operates with a 10-dB fade margin to allow for shadowing of handheld units. The network capacity is about 13,750 telephone channels. Quaternary phase-shift keying modulation is used, with *frequency-division multiple access* (FDMA)/*time-division multiple access* (TDMA). Dual-mode handsets are used that can be switched between GSM mode and satellite mode. Service features include voice telephony, fax, data, short messaging, location determination, emergency services, and high-power alerting. The mobile links operate at frequencies in the L-band (uplink 1625.5–1660.5 MHz; downlink 1525–1559 MHz), and the gateway frequencies are in the C band (uplink 6.425–6.725 GHz; downlink 3.400–3.700 GHz).

Further information can be obtained from the Web site at http://www.thuraya.com/

MSAT. This system has two geostationary satellites, MSAT-1 and MSAT-2, which provide services across North and Central America, northern South America, the Caribbean, Hawaii, and in coastal waters. The system is operated by *Mobile Satellite Ventures* (MSV), which has

offices in Reston, VA, and in Ottawa, Ontario, Canada. A variety of services are offered, including tracking and managing trucking fleets, wireless phone, data and fax, dispatch radio services, and differential GPS. L-band frequencies are used for the satellite services, the downlink band being 1530 to 1559 MHz, and the uplink, 1631.5 to 1660.5 MHz. The feeder links to the ground segment (which connects with public and private telephone and data networks) are in the Ku band, the downlink band being 10.75 to 10.95 GHz and the uplink 13.0 to 13.15 GHz and also 13.2 to 13.25 GHz. Two transponders are used for the Ku-band to L-band forward link, and one transponder is used for the L-band to Ku-band return link. Power output for the L-band is 600 W. Two L-band mesh reflector antennas, 5.7m $\times$ 4.7m are used to give an EIRP of 57.3 dBW at the edge of the service area. Power output for the Ku band is 110 W. Satellite operation and maintenance is carried out by Telesat Canada. Further information can be obtained from the Web site at http://www.msvlp.com/

17.3 VSATs

VSAT stands for *very small aperture terminal* system. This is the distinguishing feature of a VSAT system, the earth-station antennas being typically less than 2.4 m in diameter (Rana et al., 1990). The trend is toward even smaller dishes, not more than 1.5 m in diameter (Hughes et al., 1993). In this sense, the small TVRO terminals described in Sec. 16.9 for direct broadcast satellites could be labeled as VSATs, but the appellation is usually reserved for private networks, mostly providing two-way communications facilities. Typical user groups include banking and financial institutions, airline and hotel booking agencies, and large retail stores with geographically dispersed outlets.

The basic structure of a VSAT network consists of a hub station which provides a broadcast facility to all the VSATs in the network and the VSATs themselves which access the satellite in some form of multiple-access mode. The hub station is operated by the service provider, and it may be shared among a number of users, but of course, each user organization has exclusive access to its own VSAT network. Time division multiplex is the normal downlink mode of transmission from hub to the VSATs, and the transmission can be broadcast for reception by all the VSATs in a network, or address coding can be used to direct messages to selected VSATs.

Access the other way, from the VSATs to the hub, is more complicated, and a number of different methods are in use, many of them proprietary. A comprehensive summary of methods is given in Rana et al. (1990). The most popular access method is FDMA, which allows the use of comparatively low-power VSAT terminals (see Sec. 14.7.12). TDMA also can be

used but is not efficient for low-density uplink traffic from the VSAT. The traffic in a VSAT network is mostly data transfer of a bursty nature, examples being inventory control, credit verification, and reservation requests occurring at random and possibly infrequent intervals, so allocation of time slots in the normal TDMA mode can lead to low channel occupancy. A form of *demand assigned multiple access* (DAMA) is employed in some systems in which channel capacity is assigned in response to the fluctuating demands of the VSATs in the network. DAMA can be used with FDMA as well as TDMA, but the disadvantage of the method is that a *reserve channel* must be instituted through which the VSATs can make requests for channel allocation. As pointed out by Abramson (1990), the problem of access then shifts to how the users may access the reserve channel in an efficient and equitable manner. Abramson presents a method of *code-division multiple access* (CDMA) using spread spectrum techniques, coupled with the Aloha protocol. The basic Aloha method is a random-access method in which packets are transmitted at random in defined time slots. The system is used where the packet time is small compared with the slot time, and provision is made for dealing with packet collisions which can occur with packets sent up from different VSATs. Abramson calls this method *spread Aloha* and presents theoretical results which show that the method provides the highest throughput for small earth stations.

VSAT systems operate in a star configuration, which means that the connection of one VSAT to another must be made through the hub. This requires a double-hop circuit with a consequent increase in propagation delay, and twice the necessary satellite capacity is required compared with a single-hop circuit (Hughes et al., 1993). In Hughes, a proposal is presented for a VSAT system which provides for *mesh connection*, where the VSATs can connect with one another through the satellite in a single hop.

Most VSAT systems operate in the Ku band, although there are some C-band systems in existence (Rana et al., 1990). For fixed-area coverage by the satellite beam, the system performance is essentially independent of the carrier frequency. For fixed-area coverage, the beamwidth and hence the ratio λ/D is a constant (see Eq. 6.33). The satellite antenna gain is therefore constant (see Eq. 6.32), and for a given high-power amplifier output, the satellite EIRP remains constant. As shown in Sec. 12.3.1, for a given size of antenna at the earth station and a fixed EIRP from the satellite, the received power at the earth station is independent of frequency. This ignores the propagation margins needed to combat atmospheric and rain attenuation. As shown in Hughes et al. (1993), the necessary fade margins are not excessive for a Ka-band VSAT system, and the performance otherwise is comparable with a Ku-band system. (From Table 1.1, the K band covers 18 to 27 GHz and Ka band covers

27 to 40 GHz. In Hughes (1993), results are presented for frequencies of 18.7 and 28.5 GHz.)

As summarized in Rana et al. (1990), the major shortcomings of present-day VSAT systems are the high initial costs, the tendency toward optimizing systems for large networks (typically more than 500 VSATs), and the lack of direct VSAT-to-VSAT links. Technological improvements, especially in the areas of microwave technology and digital signal processing (Hughes et al., 1993), will result in VSAT systems in which most, if not all, of these shortcomings will be overcome.

17.4 Radarsat

Radarsat is an earth-resources remote-sensing satellite, which is part of the Canadian space program. Radarsat-1 was launched on November 4, 1995, and Radarsat-2 is scheduled for launch in 2006. The objectives of the Radarsat program, as stated by the Canadian Space Agency, are to:

- Provide application benefits for resource management and maritime safety
- Develop, launch, and operate an earth observation satellite with *synthetic aperture radar* (SAR). Establish a Canadian mission control facility
- Market Radarsat data globally through a commercial distributor
- Make SAR data available for research
- Map the whole world with stereo radar
- Map Antarctica in two seasons

The applications seen for Radarsat are:

- Shipping and fisheries
- Ocean feature mapping
- Oil pollution monitoring
- Sea ice mapping (including dynamics)
- Iceberg detection
- Crop monitoring
- Forest management
- Geological mapping (including stereo SAR)
- Topographic mapping
- Land use mapping

The Radarsat satellites are planned to fly in a low-earth near-circular orbit. The orbital details are given in Table 17.1.

It will be seen that the orbital parameters are similar to those for the *National Oceanographic and Atmospheric Administration* (NOAA) satellites described in Chap. 1. In particular, the Radarsat orbit is sun synchronous. There are fundamental differences, however. Radarsat carries only C-band radar as the sensing mechanism, whereas the NOAA satellites carry a wide variety of instruments, as described in Secs. 1.5 and 7.11. Even though it is known that C-band radar is not the optimal sensing mechanism for all the applications listed, the rationale for selecting it is that it does penetrate cloud cover, smoke, and haze, and it does operate in darkness. Much of the sensing is required at high latitudes, where solar illumination of the earth can be poor and where there can be persistent cloud cover.

It also will be seen that the orbit is described as *dawn to dusk*. What this means is that the satellite is in view of the sun for the ascending and descending passages. With the radar sensor it is not necessary to have the earth illuminated under the satellite; in other words, the sun's rays reach the orbital plane in a broadside fashion. The main operational advantage, suggested in Raney et al. (1991), is that the radar becomes fully dependent on solar power rather than battery power for both the ascending and descending passes. Since there is no operational need to distinguish between the ascending and descending passes, nearly twice as many observations can be made than otherwise would be possible. Also, as Raney et al. point out, the downlink periods for data transmission from Radarsat will take place at times well removed from those used by other remote-sensing satellites. Further advantages stated by the Canadian Space Agency are that the solar arrays do not have to rotate, the arrangement leads to a more stable thermal design for the spacecraft, the spacecraft design is simpler, and it provides for better power-raising capabilities. With this particular dawn-to-dusk orbit, the satellite will be eclipsed by the earth in the southern hemisphere from May 15 to July 30. The eclipse period changes gradually from zero to a maximum of about 15 min and back again to zero, as shown in Fig. 17.1. The battery backup consists of three 50 Ah nickel-cadmium batteries.

TABLE 17.1 Radarsat Orbital Parameters

Geometry	Circular, sun synchronous (dawn–dusk)
Altitude (local)	798 km
Inclination	98.6°
Period	100.7 min
Repeat cycle	24 days

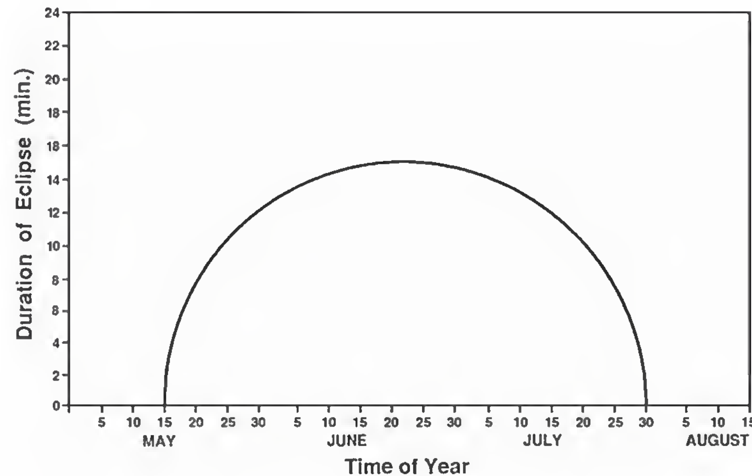


Figure 17.1 Duration of eclipse versus time of year, dawn–dusk orbit. (*Courtesy of Canadian Space Agency.*)

Radarsat, shown in Fig. 17.2, is a comparatively large spacecraft, the total mass in orbit being about 3100 kg. The radar works at a carrier frequency of 5.3 GHz, which can be modulated with three different pulse widths, depending on resolution requirements. The SAR operating modes are illustrated in Fig. 17.3. The swath illuminated by the radar lies 20° to the east and parallel to the subsatellite path. As shown, different beam configurations can be achieved, giving different resolutions. The mode characteristics are summarized in Table 17.2.

The satellite completes 14 and 7/24 revolutions per day. The separation between equatorial crossings is 116.8 km. According to Raney et al. (1991), the scanning SAR is the first implementation of a special radar technique.

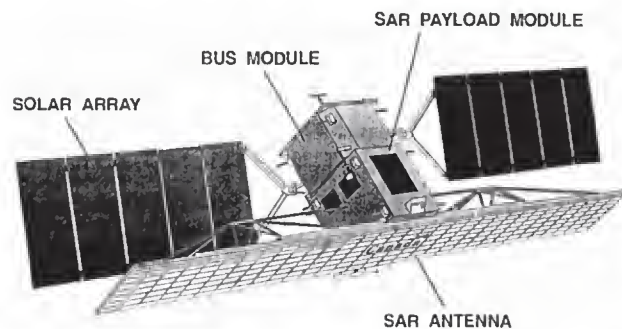


Figure 17.2 Spacecraft configuration. (*Courtesy of Canadian Space Agency.*)

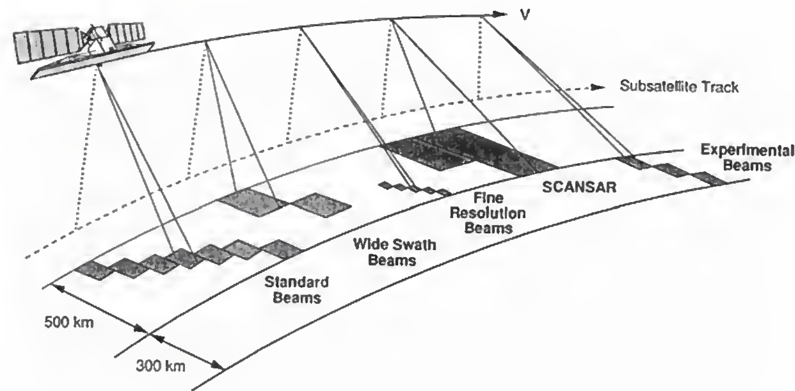


Figure 17.3 SAR operating modes. (*Courtesy of Canadian Space Agency.*)

Radarsat-2 is scheduled for launch in 2006. This will have the same orbit and repeat cycle as Radarsat-1, and will provide a follow-on program for all the Radarsat-1 imaging modes but with additional capabilities. The radar resolution with Radarsat-2 is 2.5 to 4.5 m compared to over 9 m with Radarsat-1. Whereas Radarsat-1 uses horizontal (H) polarization only, Radarsat-2 will be able to transmit and receive horizontal and vertical (V) in four modes: HH; HV; VH; and VV. Interpretation of radar data is greatly improved through the use of multipolarization. More detailed information can be obtained from http://www.radarsat2.info/

17.5 Global Positioning Satellite System (GPS)

In the GPS system, a constellation of 24 satellites circles the earth in near-circular inclined orbits. By receiving signals from at least four of these satellites, the receiver position (latitude, longitude, and altitude) can be determined accurately. In effect, the satellites substitute for the geodetic position markers used in terrestrial surveying. In terrestrial

TABLE 17.2 SAR Modes

	Swath, km	Resolution	Incidence angle, degrees
Operational	1	28 m × 30 m (4 looks)	20–50
High resolution	50	10 m × 10 m (1 look)	30–50
Experimental	100	28 m × 30 m (4 looks)	50–60
Scan SAR	500	100 m × 100 m (6 looks)	20–50

surveying, it is only necessary to have three such markers to determine the three unknowns of latitude, longitude, and altitude by means of triangulation. With the GPS system, a time marker is also required, which necessitates getting simultaneous measurements from four satellites.

The GPS system uses one-way transmissions, from satellites to users, so that the user does not require a transmitter, only a GPS receiver. The only quantity the receiver has to be able to measure is time, from which propagation delay, and hence the range to each satellite, can be determined. Each satellite broadcasts its ephemeris (which is a table of the orbital elements as described in Chap. 2), from which its position can be calculated. Knowing the range to three of the satellites and their positions, it is possible to compute the position of the observer (user). The geocentric-equatorial coordinate system (see Sec. 2.9.6) is used with the GPS system, where it is called the *earth-centered, earth-fixed* (ECEF) coordinate system. Denoting the coordinates for satellite n by (x_n, y_n, z_n) and those for the observer by (x_0, y_0, z_0) the range from observer to satellite ρ_{On} is obtained from

$$\rho_{On}^2 = (x_n - x_0)^2 + (y_n - y_0)^2 + (z_n - z_0)^2 \tag{17.1}$$

As Eq. (17.1) shows, there are three unknowns (x_0, y_0, z_0) and in theory only three equations (for $n = 1$, 2, and 3) are needed. In practice, measurements for four satellites must be made (n ranges from 1 to 4) and this will be explained shortly. For the moment, knowing the positions of three satellites and the measured range to each, the position of the observer relative to the coordinate system can be calculated. Of course, the satellites are moving, so their positions must be tracked accurately. The satellite orbits can be predicted from the orbital parameters (as described in Chap. 2). These parameters are continually updated by a master control station which transmits them up to the satellites, where they are broadcast as part of the navigational message from each satellite.

Just as in a land-based system, better accuracy is obtained by using reference points well separated in space. For example, the range measurements made to three reference points clustered together will yield nearly equal values. Position calculations involve range differences, and where the ranges are nearly equal, any error is greatly magnified in the difference. This effect, brought about as a result of the satellite geometry, is known as *dilution of precision* (DOP). This means that range errors occurring from other causes, such as timing errors, are magnified by the geometric effect. With the GPS system, dilution of position is taken into account through a factor known as the *position dilution of precision* (PDOP) *factor*. This is the factor by which the range errors are multiplied

to get the position error. The GPS system has been designed to keep the PDOP factor less than 6 most of the time (Langley, 1991*c*).

The GPS constellation consists of 24 satellites in six near-circular orbits, at an altitude of approximately 20,000 km (Daly, 1993). The ascending nodes of the orbits are separated by 60°, and the inclination of each orbit is 55°. The four satellites in each orbit are irregularly spaced to keep the PDOP factor within the limits referred to earlier.

Time enters into the position determination in two ways. First, the ephemerides must be associated with a particular time or epoch (as described in Chap. 2). The standard timekeeper is an atomic standard, maintained at the U.S. Naval Observatory, and the resulting time is known as *GPS time*. Each satellite carries its own atomic clock. The time broadcasts from the satellites are monitored by the control station, which transmits back to the satellites any errors detected in timing relative to GPS time. No attempt is made to correct the clocks aboard the satellites; rather, the error information is rebroadcast to the user stations, where corrections can be implemented in the calculations. It can be assumed, therefore, that the satellite position relative to the ECEF coordinate system (the geocentric-equatorial coordinate system, Sec. 2.9.6), is accurately known. The coordinates for satellite n will be denoted by (x_n, y_n, z_n).

Second, time markers are needed to show when transmissions leave the satellites so that, by measuring propagation times and knowing the speed of propagation, the ranges can be calculated. Therein lies a problem, since the user stations have no direct way of telling when a transmission from a satellite commenced. The problem is overcome by having the satellite transmit a continuous-wave carrier, which is modulated by a pseudo-random code, timing for the carrier and the code being derived from the atomic clock aboard the satellite. At a user station, the receiver generates a replica of the modulated signal from its own (nonatomic) clock, which is correlated with the received signal in a correlator. A delay is introduced into the replica signal path and is adjusted until the two signals show maximum correlation. If the receiver clock started at exactly the same time as the satellite clock, the delay in the replica path would be equal to the propagation delay. As it is there will be some unknown difference in the start times. Let t_n be the true propagation time from satellite n to the observer, and let t_{dn} be the delay time as measured by the correlator. Let Δt be the unknown difference between the clock start times then $t_n = t_{\text{dn}} - \Delta t$ (Δt can be + or –). The satellite clocks are synchronized to the master atomic clock, so Δt will be the same for all satellites. The range to the satellite n is therefore

$$\rho_{On} = ct_n = c(t_{\text{dn}} - \Delta t) \tag{17.2}$$

where c is the speed of light. Substituting this in Eq. (17.1) gives:

$$O = (x_n - x_0)^2 + (y_n - y_0)^2 + (z_n - z_0)^2 - c^2(t_{dn} - \Delta t)^2 \qquad (17.3)$$

The unknowns here are the location (x_0, y_0, z_0) and the timing difference Δt. For satellite n the position (x_n, y_n, z_n) is known and delay time t_{dn} is measured by the receiver. Since there are four unknowns, the receiver must be capable of measuring t_{dn} for four satellites simultaneously ($n = 1, 2, 3, 4$) to yield four simultaneous equations of the form (17.3). These can be solved to find the unknowns Δt and (x_0, y_0, z_0). The latter, of course, are the required position coordinates for the receiver, and these would be converted into local coordinates (latitude, longitude, and altitude). All this requires quite sophisticated microprocessing in the receiver. Also, the composition of the GPS signal is much more complex than indicated here, utilizing spread-spectrum techniques.

The free-space value for c is used where high precision is not required. However, the free-space value cannot be used for radio waves traveling through the ionosphere and the troposphere. Although the change in propagation velocity is small in absolute terms, it can introduce significant timing errors in certain applications. Also, the satellites clocks, although highly accurate will have their own timing errors. The dilution of position errors, described previously, combined with these timing errors set a limit on the accuracy of location determination. Where very high accuracy is required differential GPS (DGPS) can be used. Two receivers are used, one of which is placed at an accurately known location. Thus, the reference receiver makes a measurement in the usual way, but since the coordinates for the location are known, the only unknown in Eq. (17.1) is the range ρ_{On}, which can now be calculated. Comparing the reference value with the value obtained from the receiver correlator enables the errors to be determined. The reference receiver is linked by radio to the receiver at the unknown location, which can now correct for the errors. The two receivers may be up to a few hundred kilometers apart but this is insignificant in comparison with the distances to the satellites, and it may be assumed also that the signal paths through the ionosphere and troposphere are the same.

Further details of the GPS system will be found in: Langley (1990*a*, 1991*b*, 1991*c*), Kleusberg and Langley (1990), and Mattos (1992, 1993*a*, 1993*b*, 1993*c*, 1993*d*, 1993*e*) and at http://www.trimble.com/gps/why.html

17.6 Orbcomm

The *Orbital Communications Corporation* (Orbcomm) system is a LEO satellite system, which provides two-way message and data communications services and position determination. The satellite constellations are shown in Table 17.3

TABLE 17.3 Orbital Parameters for ORBCOMM

Orbital plane	A	B	C	D	E	F
Launch date	12/23/97	8/2/98	9/23/98	12/4/99	4/3/95	2/10/98
No. of Satellites	8	8	6	6	1	1
Semimajor axis, km	7.178	7.178	7.178	7.178	7.078	7.178
Altitude, km	800	800	800	800	710	820
Inclination, deg	45	45	45	45	70	108
Period, min	101	101	101	101	99	101

SOURCE: http://www.orbcomm.com/wwwroot/

A number of *gateway control centers* (GCC) in various countries provide the switching necessary to link subscribers with terrestrial networks. The GCCs also provide a performance and status monitoring service for the system. The *network control center* (NCC) located in Dulles, VA serves as the U.S. GCC and also manages the satellite constellation. *Gateway earth stations* (GES) throughout the world link the ground segment with the satellites. Users have a subscriber communicator for fixed and mobile messaging. Figure 17.4 illustrates the system and Table 17.4 shows some of the parameters used in link-budget calculations.

The satellites are small compared with the geostationary satellites in use, as shown in Fig. 17.5. The VHF/UHF antennas are seen to extend in a lengthwise manner, with the solar panels opening like lids top and bottom. Before launch, the satellites are in the shape of a disk, and the launch vehicle, a Pegasus XL space booster [developed by *Orbital Sciences Corporation* (OSC), the parent company of Orbcomm] can deploy eight satellites at a time into the same orbital plane. For launch, the satellites are stacked like a roll of coins, in what the company refers to as "an eight-pack."

Attitude control is required to keep the antennas pointing downward and at the same time to keep the solar panels in sunlight (battery backup is provided for eclipse periods). A three-axis magnetic control system, which makes use of the earth's magnetic field, and gravity gradient stabilization are employed. A small weight is added at the end of the antenna extension to assist in the gravity stabilization. Thus, the satellite antennas hang down as depicted in Fig. 17.4. At launch, the initial separation velocity is provided by springs used to separate the satellites, and a braking maneuver is used when the satellites reach their specified 45° in-plane separation. Intraplane spacing is maintained by a proprietary station-keeping technique which, it is claimed, has no cost in terms of fuel usage (Orbcomm, 1993). Because no onboard fuel is required to maintain the intraplane spacing between satellites, the satellites have a design lifetime of 4 years, which is based on the projected degradation of the power subsystem (solar panels and batteries).

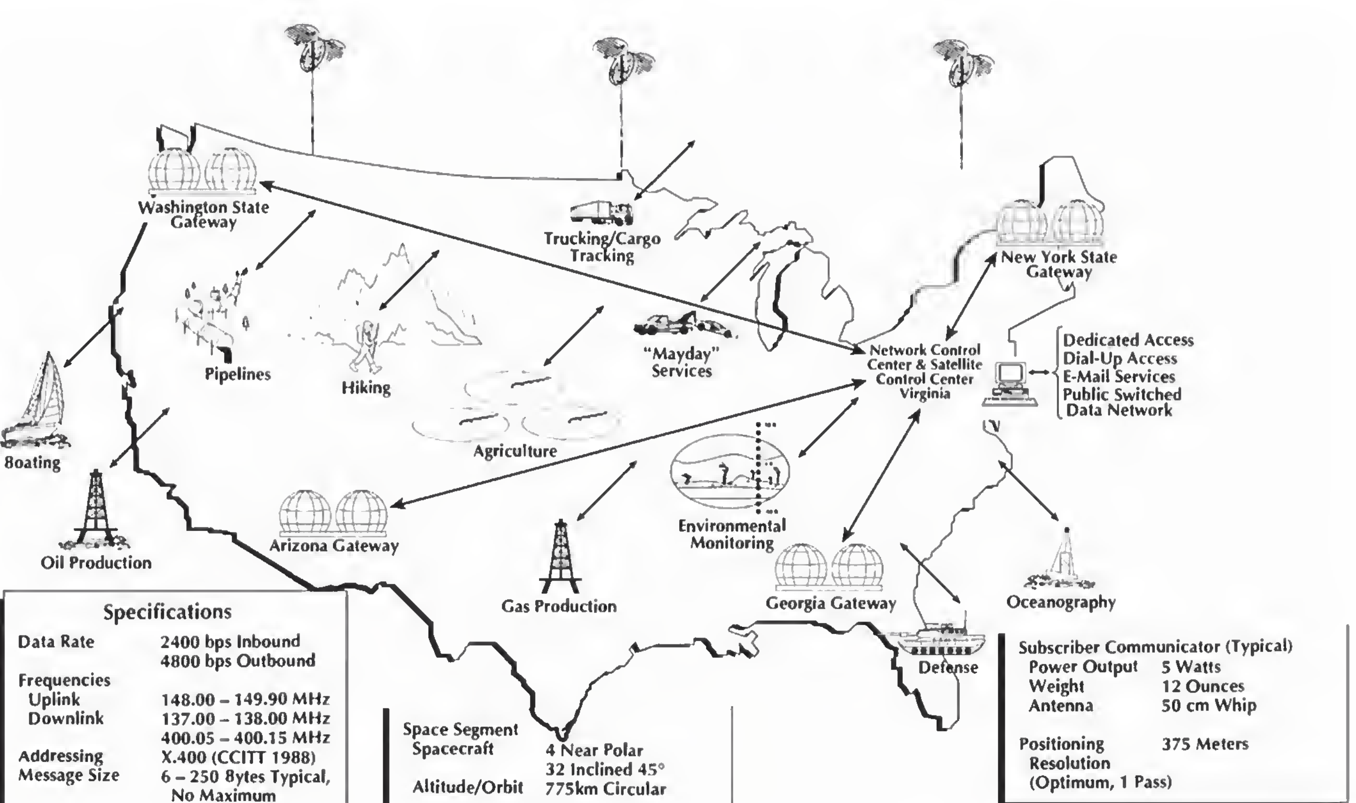


Figure 17.4 The Orbcomm system. (*Courtesy of Orbital Communications Corporation.*)

TABLE 17.4 Link-Budget Parameters

Range 2730 km	Gateway-satellite Uplink	Gateway-satellite Downlink	Satellite-subscriber Uplink	Satellite-subscriber Downlink
Frequency, MHz	149.4	137.2	148.95	137.5
[EIRP], dBW	40	6.5	7.5	12.5
Misc. losses, dB	7.3	7.3	11.3	11.3
[G/T], dBK^{-1}	−33.3	−12.8	−26	−28.6
Data rate, kbps	57.6	57.6	2.4	4.8

SOURCE: Orbcomm, 1993.

The messaging and data channels are located in the VHF band, the satellites receiving in the 148 to 149.9-MHz band and transmitting in the 137 to 138-MHz band. Circular polarization is used. In planning the frequency assignments, great care has been taken to avoid interference to and from other services in the VHF bands; the reader is referred to Orbcomm (1993) for details. In particular, the subscriber-to-satellite uplink channels utilize what is termed a *dynamic channel activity*

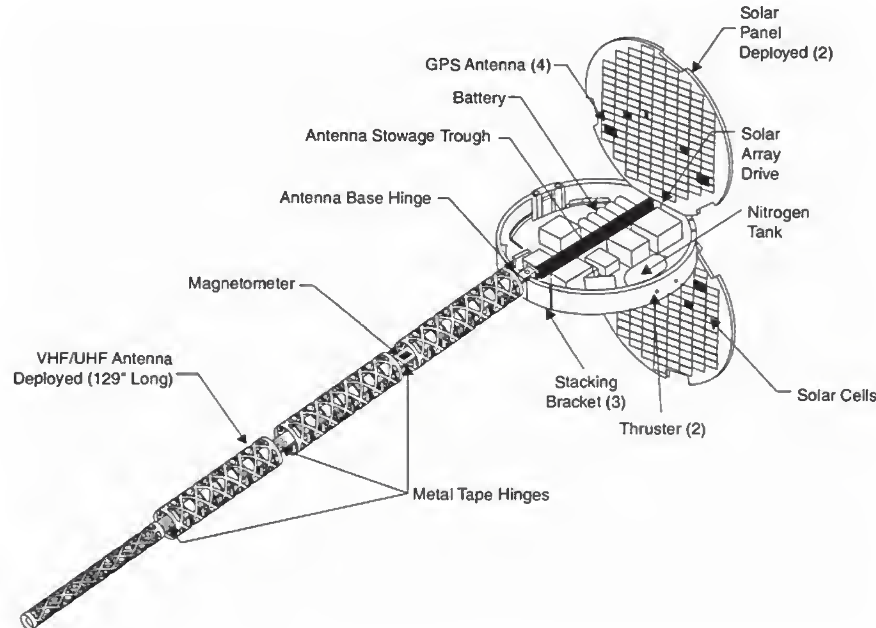


Figure 17.5 Orbcomm/Microstar satellite. (*Courtesy of Orbital Communications Corporation.*)

assignment system (DCAAS), in which a scanning receiver aboard the satellite measures the interference received in small bandwidths, scanning the entire band every 5 s or less. The satellite receiver can then prepare a list of available channels (out of a total of 760) and prioritize these according to interference levels expected.

The Orbcomm system is capable of providing subscribers with a basic position determination service through the use of Doppler positioning, which fixes position to within a few hundred meters. The beacon signal at 400.1 MHz is used to correct for errors in timing measurements introduced by the ionosphere (these errors are also present in the GPS system described in Sec. 17.5, and two frequencies are used in that situation for correction purposes). When used in conjunction with the VHF downlink signal, the beacon signal enables the effects of the ionosphere to be removed. It will be observed from Fig. 17.5 that the satellites carry GPS antennas, which enable onboard determinations of the positions of the satellites. This information can then be downloaded on the VHF subscriber channel and used for accurate positioning.

One significant advantage achieved with LEOs is that the range is small compared with geostationary satellites (the altitude of the Orbcomm satellites is typically 800 km compared with 35,876 km for geostationary satellites). Thus, the *free-space loss* (FSL) is very much less. Propagation delay is correspondingly reduced, but this is not a significant factor where messaging and data communications, as compared to real-time voice communications, are involved.

The Orbcomm system provides a capacity of more than 60,000 messages per hour. By using digital packet switching technology, and confining the system to nonvoice, low-speed alphanumeric transmissions, Orbcomm calculates that the service, combined with other LEO systems, will be able to provide for 10,000 to 20,000 subscribers per kilohertz of bandwidth, which is probably unmatched by any other two-way communications service. Although it is a U.S.-based system, because of the global nature of satellite communications, Orbcomm has signed preliminary agreements with companies in Canada, Russia, South Africa, and Nigeria to expand the Orbcomm service (Orbcomm, 1994). The Orbcomm Web site is http:www.orbcomm.com

17.7 Iridium

The Iridium concept was originated by engineers at Motorola's Satellite Communications Division in 1987. Originally envisioned as consisting of 77 satellites in LEO, the name Iridium was adopted by analogy with the element *Iridium* which has 77 orbital electrons. Further studies led to a revised constellation plan requiring 66 satellites (Leopold, 1992). The original company, Iridium LLC went bankrupt in 2000. In December

of that year a group of private investors organized Iridium Satellite LLC, which acquired the operating assets of the bankrupt Iridium LLC including the satellite constellation, the terrestrial network, Iridium real property, and intellectual capital.

The 66 satellites are grouped in 6 orbital planes each containing 11 active satellites. The orbits are circular, at a height of 780 km (485 miles). Prograde orbits are used, the inclination being 86.4°. The eleven satellites in any given plane are uniformly spaced, the nominal spacing being 32.7°. An in-orbit spare is available for each plane at 130 km lower in the orbital plane. Some of the other orbital characteristics are listed in Table 17.5.

The satellites travel in corotating planes, that is, they travel up one side of the earth, cross over near the north pole, and travel down the other side. Keeping in mind that there are eleven equispaced satellites in each plane, it will be seen that both sides of the earth are covered continuously. The satellites in adjacent planes travel out of phase, meaning that adjacent planes are rotated by half the satellite spacing relative to one-another. This is sketched in Fig.17.6, which shows the view from above the north pole. Collision avoidance is built into the orbital planning, and the closest approach between satellites is 223 km. Satellites in planes 1, 3, and 5 cross the equator in synchronization, while satellites in planes 2, 4, and 6 also cross in synchronization, but out of phase with those in planes 1, 3, and 5.

Although the planes are corotating, the first and last planes must be counter-rotating where they are adjacent. This is illustrated Fig. 17.6. The separation between corotating planes is 31.6° which allows 22° separation between the first and last planes. The closer separation is needed

TABLE 17.5 Physical Characteristics of Satellite

Characteristic	Nominal values (3 sigma range for km and degrees)
Station keeping	2.0 km cross-track 5.7 km in-track 4.7 km radial
Antenna pointing accuracy toward earth	1.0° in azimuth 0.7° in elevation
Attitude stabilization and station-keeping systems	Roll: <0.5° Pitch: <0.4° Yaw: <0.75°
Electrical energy system	<1200 W solar array <50 A-h battery
Estimated minimum lifetime of in-orbit satellite	5 years

SOURCE: Motorola, 1992.

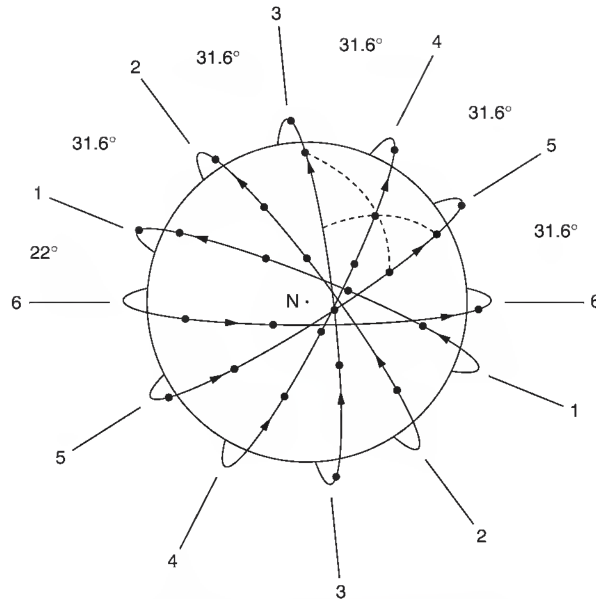


Figure 17.6 A polar view of the Iridium satellite orbits.

because earth coverage under the counter-rotating "seam" is not as efficient as it is under the corotating seams. Two-way communication links exist between each satellite and its nearest neighbors ahead and front, and to the satellites in the adjacent planes, as shown by the dotted lines in Fig. 17.6.

Figure 17.7 shows a system overview. The up/down links between subscribers and satellites take place in the L-band. A 48-beam antenna pattern is used from each satellite with each beam under separate control. At the equator, for instance, overlap of patterns will be minimal and all beams may be on, while at high latitudes, considerable overlap occurs, and certain beams will be switched off. Also, in regions where operation is prohibited by the telecommunications administration, the beams can be switched off. The switching of beams is referred to as *cell management*. The beams are similar to the cells encountered in cellular mobile, but with the fundamental difference that the beams move relative to the subscriber, whereas in cellular mobile, the cells are fixed. It must be realized that the satellites are traveling at a much greater velocity than that normally encountered with terrestrial vehicles, which may be considered stationary relative to the satellites. The orbital period is approximately 100 min. Using an average value of 6371 km for the earth's radius the surface speed is $(2 \times 6371 \times \pi)/100 \approx 400$ km/min or just over 15000 mph. The 48 cell pattern is shown in Fig. 17.8.

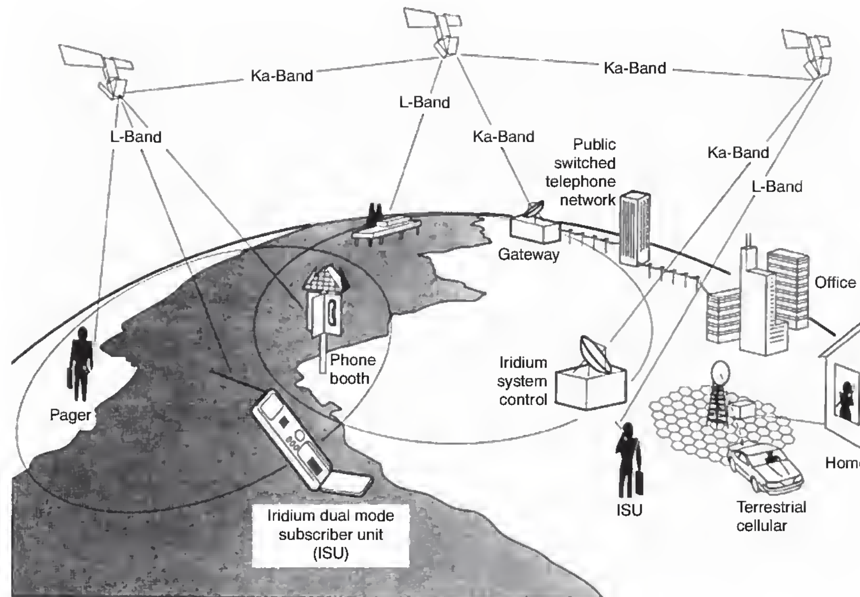


Figure 17.7 Iridium system overview. (*Courtesy of Motorola.*)

Intersatellite links, and the up/down links between satellites and gateway stations operate in the Ka band. The link between the Iridium system control station and the satellites is also in the Ka band. Circular polarization is used on all links (Leopold 1992). The multiple access utilizes a combination of TDMA and FDMA. Each subscriber unit operates in a burst-mode using a single carrier transmission (Motorola 1990). The TDMA frame format is shown in Fig. 17.9 and the uplink/downlink RF frequency plan in Fig. 17.10 (Motorola 1992). Fig. 17.11 shows an artist's conception of satellites to be used.

Table 17.6 shows link values for the *space vehicle* (SV) to *Iridium Subscriber Unit* (ISU) for cells 1 and 16 of Fig. 17.8. In this table, *shadowing* refers to the added attenuation resulting from natural vegetation. Tables 17.7 and 17.8 show the other link values.

The company stress that Iridium is not a replacement for existing cellular systems, but rather an extension of wireless telephony. The major advantages listed by the company are:

- Iridium is not likely to be disabled in disaster situations (earthquakes, fire, floods and the like), and will therefore be available as an emergency service in the event that terrestrial cellular services are knocked-out.
- Iridium is able to provide mobile services to those areas (e.g, remote, and sparsely populated areas) that are not reached by terrestrial cellular services.

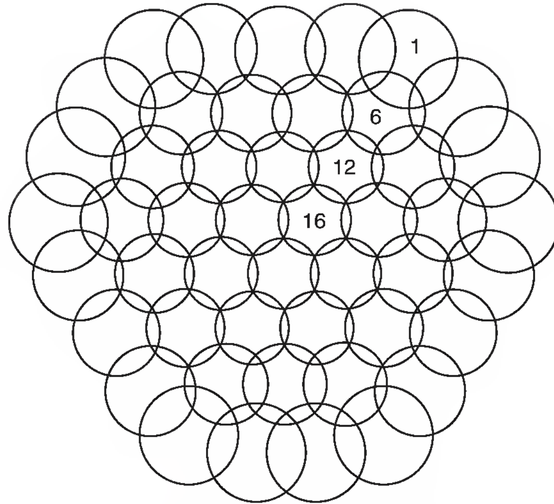


Figure 17.8 48-cell L-band integrated antenna pattern architecture. (*Courtesy of Motorola 1992.*)

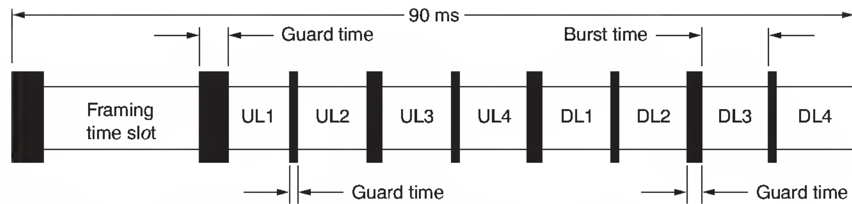


Figure 17.9 TDMA frame format. (*Courtesy of Motorola 1992.*)

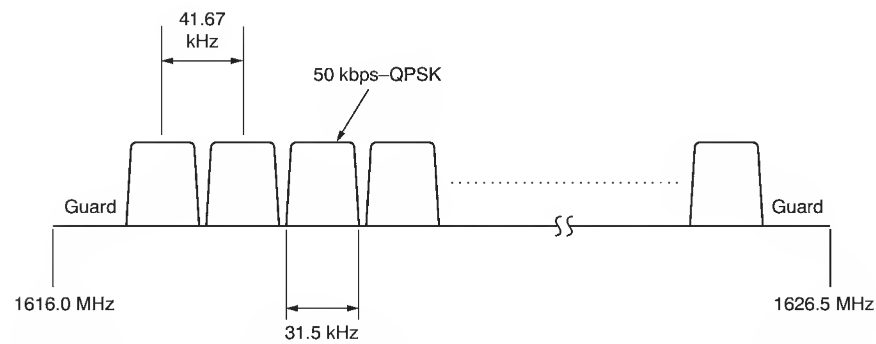


Figure 17.10 L-band uplink/downlink RF frequency plan. (*Courtesy of Motorola 1992.*)

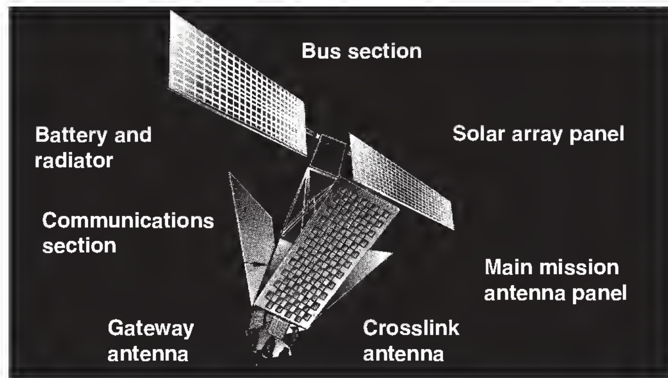


Figure 17.11 Motorola SV (Space Vehicle) deployed configuration. (*Courtesy of Motorola.*)

TABLE 17.6 SV–ISU Link Values

	Cell-1 slant range 2461.7 km		Cell-16 slant range 960 km	
	Uplink	Downlink	Uplink	Downlink
[EIRP], dBW	−4.2 (6)	15.7 (27.7)	−4.9 (6)	7.5 (19.5)
[G_R], dB	23.9	1	16.4	1
T_S K	500	250	500	250

NOTES: Frequency: 1621.25 MHz
Miscellaneous propagation losses: no shadowing 0.7 dB; with shadowing 15.7 dB
Coded data rate: 50 kbps
The EIRP values adjusted for shadowing are shown in parenthesis.

TABLE 17.7 SV–Gateway Link Values

	Downlink–20 GHz		Uplink–29.40 GHz	
	Rain	Clear	Rain	Clear
[EIRP], dBW	23.2	13.5	68	43.2
Misc. losses, dB	17.8	5.1	33.1	4.6
[G/T], dBK^{-1}	24.6	24.6	−1.02	−1.02

NOTES: Range: 2326 km
Coded data rate: 6.25 Mbps

TABLE 17.8 SV–SV

Frequency: 23.28 GHz
[EIRP]: 38.4 dBW
Coded data rate: 25 Mbps
Receive antenna gain: $[G_R]$: 36.7 dB
Receive system noise temperature: 720.3 K (1188.3 K)
Range: East–West 4400 km; North–South 4050 km
Miscellaneous losses: East–West 3.6 dB (1.8 dB); North–South 4.4 dB (1.8 dB).

NOTES: Values "with sun" are shown in parenthesis. The miscellaneous losses include a margin which is omitted for the "with sun" situation.

- Iridium is able to offer more channels with shorter delays, and the ability to connect into worldwide networks for those areas which currently obtain mobile services through geostationary satellites.
- Iridium does make it possible to get a service in operation very quickly in areas where no telephone services exist, while terrestrial networks are being installed.

17.8 Problems and Exercises

17.1. Write brief notes on the advantages and disadvantages of using satellites in LEOs, MEOs, and GEOs for mobile satellite communications.

17.2. Write brief notes on the advantages and disadvantages of onboard switching and routing compared to the "bent pipe" mode of operation for satellite mobile communications.

17.3. Describe the operation of a typical VSAT system. State briefly where VSAT systems find widest application.

17.4. Describe the main features of Radarsat. Explain what is meant by a "dawn to dusk" orbit and why the Radarsat follows such an orbit.

17.5. Explain why a minimum of four satellites must be visible at an earth location utilizing the GPS system for position determination. What does the term *dilution of position* refer to?

17.6. Describe the main features and services offered by the Orbcomm satellite system. How do these services compare with services offered by geostationary satellites and terrestrial cellular systems?

17.7. Calculate the free-space loss for the GES to satellite uplink in the Orbcomm system.

17.8. Calculate the received $[C/N_0]$ value and the received $[E_b/N_0]$ for the GES to satellite uplink in the Orbcomm system.

17.9. Repeat Prob. 17.8 calculations for the satellite to GES downlink.

17.10. Calculate the received $[C/N_0]$ value and the received $[E_b/N_0]$ for the subscriber to satellite uplink in the Orbcomm system.

17.11. Repeat Prob. 17.10 calculations for the satellite to subscriber downlink in the Orbcomm system.

17.12. Describe the main features of the Iridium system, and comment briefly on how this differs from the Orbcomm system.

17.13. For Iridium, calculate the uplink values of $[C/N_0]$ for cell-1 and cell-16 under no shadowing conditions. Calculate the corresponding $[E_b/N_0]$ values.

17.14. Repeat the calculations in problem 17.13 for the downlink.

17.15. For the shadowing situation with Iridium, calculate the uplink values of $[C/N_0]$ for cell-1 and cell-16. Calculate the corresponding $[E_b/N_0]$ values.

17.16. Repeat the calculations in Prob. 17.15 for the downlink.

17.17. Calculate for the Iridium SV–Gateway downlink the $[E_b/N_0]$ values for (*a*) rain conditions and (*b*) clear conditions.

17.18. Repeat the calculations of Prob. 17.17 for the Iridium SV–Gateway uplink.

17.19. Calculate for the Iridium east-west intersatellite link the $[C/N_0]$ values for (*a*) no sun, and (*b*) sun conditions.

17.20. Repeat the calculations of Prob. 17.19 for the north–south intersatellite link.

References

Abramson, N. 1990. "VSAT Data Networks." *Proc. IEEE*, Vol. 78, No. 7, July, pp. 1267–1274.

Clarke, A.r C. 1945. "Extraterrestrial Relays." *Wireless World*, Vol. 51, October. pp. 305–308.

Daly, P. 1993. "Navstar GPS and GLONASS: Global Satellite Navigation Systems." *Electron. Commun. Eng. J.*, Vol. 5, No. 6, December, pp. 349–357.

Hughes, C. D., C. Soprano, F. Feliciani, and M. Tomlinson, 1993. "Satellites Systems in a VSAT Environment." *Elect. Comm. Engr. J.*, Vol. 5, No. 5, October, pp. 285–291.

Kleusberg, A., and R. B. Langley. 1990. "The Limitations of GPS." *GPS World*, Vol. 1, No. 2, March–April, pp. 50–52.

Langley, R. B. 1991*a*. "Why ls the GPS Signal So Complex?" *GPS World*, Vol. 1, No. 3, May–June, pp. 50–59.

Langley, R. B. 1991*b*. "The GPS Receiver: An lntroduction." *GPS World*, Vol. 2, No. 1, January, pp. 50–53.

Langley, R. B. 1991*c*. "The Mathematics of GPS." *GPS World*, Vol. 2, No. 7 July–August, pp. 45–50.

Leopold, R. J. 1992. the Iridium Communication system, TUANZ'92, "Communications for Competitive Advantage," *Conference and Trade Exhibition*, Aotea Centre, Auckland, New Zealand, August 10–12.

Mattos, P. 1992. "GPS." *Electronics World + Wireless World*, December, No. 1681, pp. 982–987.

Mattos, P. 1993*a*. "GPS. 2: Receiver Architecture." *Electronics World + Wireless World*, January, No. 1682, pp. 29–32.

Mattos, P. 1993*b*. "GPS. 3: The GPS Message on the Hardware Platform." *Electronics World + Wireless World*, February, No. 1683, pp. 146–151.

Mattos, P. 1993*c*. "GPS. 4: Radio Architecture." *Electronics World + Wireless World*, March, No. 1684, pp. 210–216.

Mattos, P. 1993*d*. "GPS. 5: The Software Engine." *Electronics World + Wireless World*, April, No. 1685, pp. 296–304.

Mattos, P. 1993*e*. "GPS. 6: Applications." *Electronics World + Wireless World*, May, No. 1686, pp. 384–389.

Miller, B. 1998. "Satellites Free the Mobile Phone." *IEEE Spectrum*, Vol. 35, No. 3, March, pp. 26–35.

Motorola Communications Inc. 1990. *Application of Motorola Satellite Communications Inc. for IRIDIUM, a Low Earth Orbit Satellite System, Before the Federal Communications Commission.* Motorola Communications Inc., Washington D.C., December.

Motorola Communications Inc. 1992. Minor Amendment to Application Before the FCC to Construct and Operate a Low Earth Orbit Satellite System in the RDSS Uplink Band File No. 9-DSS-P91 (87) CSS-91-010.

ORBCOMM. 1994. News release, June.

ORBCOMM. 1993. ORBCOMM Application Amendment and Supplement. File No. 22-DSS-MP-90 (20) December (addressed to Mr. William F. Caton, Acting Secretary FCC).

Rana, H. A., J. McCoskey, and W. Check. 1990. "VSAT Technology, Trends, and Applications." *Proc. IEEE*, Vol. 78, No. 7, July, pp. 1087–1095.

Sweeting, M. N. 1992. "UoSAT Microsatellite Missions." *Electron. Commun. Eng. J.*, June, Vol. 4, No. 3, pp. 141–150.

Williamson, M. 1994. "The Growth of Microsats." *IEE Review*, May, Vol. 40, No. 3, pp. 117–120.

Answers to Selected Problems

Chapter 2

2.5. 11016 km

2.7. 70.14°

2.9. (*a*) 3.842 km/s; (*b*) 12605.85 ft/s; (*c*) 8594.89 mi/h

2.11. 42165 km; 0.00018973

2.17. (*a*) 13.994 rev/day; (*b*) 0; (*c*) −3.158 deg/day

2.19. 2451248.5 day; 2451598.0 day; 2452700.1875 day; 2455382.125 day

2.21. (*a*) Jan 3, 0 h; (*b*) July 4, 03:00 h; (*c*) Oct 27, 2 h 55 min 4.21 s; (*d*) Jan 3, 7 h 3 min 57.75 s; (e) January 31, 2 h 26 min 9.6 s

2.23. (*a*) 27027 km; (*b*) 0.74718; (*c*) 12.277 rad/day; (*d*) 12.276 rad/day; (*e*) 737.024 min; (*f*) −0.149 deg/day; (*g*) 0.007 deg/day

2.25. 0.856°W

2.27. (*a*) 54.53 rad/day; (*b*) −125.425°; (*c*) 10,509.1 km; (*d*) 36.66°S

2.29. 101.83°W

2.31. 249.8°; 84.96°; 290.7°; 289.52°; 6614.9 km

2.33. −0.707 deg/day; −2.502 deg/day; 82.778 rad/day

2.35. −0.206°; 110.063 min

2.37. (*a*) 7234.6 km; (*b*) 0° Latitude approximately.

2.39. 73.521°

2.41. (*a*) 07:00 h; (*b*) 05:00 h; (*c*) 15:00 h; (*d*) 20:00 h

Chapter 3

3.3. Satellite on same longitude as the earth station, and latitude = 76.3°.

3.5. Latitude zero; maximum possible longitudinal separation 76.3°.

3.7. 302.6°; 64.8°

3.9. 44.3°; 45°

3.15. 2.1°

3.21. $\phi_{SS} = 60.025°$; $\phi_{SSmean} = 59.989°$

3.23. 35799.86 km; 35772.87 km

Chapter 4

4.3. 0.29 dB

4.7. Horizontal 0.0067 dB; vertical 0.0055 dB

4.11. (*a*) 1.12 dB; (*b*) 1.19 dB

4.13. 0.006 dB

4.15. 0.015 dB

Chapter 5

5.3. For E_{xmax} unity and $E_{ymax} = 3E_{xmax}$, the power relationship gives E_{xmax} = 10V

5.5. 0.5 W

5.7. LH elliptical

5.9. LH elliptical

5.13. 59°

5.17. 19°

5.21. PL = 0.11 dB; XPD = 16 dB

5.25. 50.4 dB

5.27. 36.6 dB

Chapter 6

6.1. 26.78 dBW

6.3. (1.41, 0.513, 2.598)

6.5. −36.87° relative to the +*y* axis

6.7. 23.87×10^{-15} W/m^2

6.9. 51.8 dB

6.11. 1.56 dB

6.13. 7.96×10^{-6} m^2

6.19. $HPBE_H = 22°$; $HPBW_E = 25.2°$; 18.72 dB

6.25. 0.9 m

6.27. 12.76 m^2; 48.07 dB

6.33. Currents are in-phase

6.37. (*a*) 1.172 cm; (*b*) 536.3 rad/m; 307.3 deg/cm

6.39. $\Delta l = 1.13$ mm

Chapter 9

9.1. 300–3400 Hz; −14.4 dBm

9.9. $Y = 0.3R + 0.59G + 0.11B$; $Q = 0.21R - 0.52G + 0.31B$; $I = 0.6R - 0.28G - 0.32B$

9.13. 500 Hz

9.17. 152 kHz

9.21. (*a*) 1.465; (*b*) 2.322; (*c*) 7.545

9.23. (*a*) 12.6 dB; (*b*) 13.9 dB; (*c*) 5.5 dB

9.25. (*a*) 62.6 dB; (*b*) 2686.9:1

Chapter 10

10.5. (*a*) 11100001, 01100001; (*b*) 11011111, 01011111; (*c*) 11010100, 01010100; (*d*) 10011000, 00011000

10.7. (*a*) 51.7 dB; (*b*) 16

10.9. (*a*) 0.079; (*b*) 3.87×10^{-6}; (*c*) approximately 0

10.15. Approximately 1.91×10^{-4}

10.17. Approximately 1.6×10^{-4}

10.19. Approximately 1.1×10^{-6}

Chapter 11

11.3. codewords 256; datawords 128

11.5. (*a*) (0000000); (*b*) (1111111); (*c*) (0010101)

11.7. s = 001

11.9. 0.839

11.11. 3

11.17. 2.117 MHz; 1.059 MHz

11.19. 2.78×10^{-4}

11.21. −3 dB

11.23. For 111, the Euclidean metric is 3.942, and for 000, it is 4.143. This is closest to 111; therefore, the soft decision decoding produces a binary 1. For hard decision, majority vote results in a binary 0.

11.25. 0.201

Chapter 12

12.1. (*a*) 14.6 dB; (*b*) 23.6 dBW; (*c*) 75.6 dBHz; (*d*) 28.2 dB; (*e*) 23 dBK

12.3. (*a*) 42.9 dB; (*b*) 50.3 dB

12.5. 1.98 m^2

12.7. (*a*) 4.33 pW/m^2; (*b*) 195.8 pW

12.9. −98.6 dBW or 138.6 pW

12.11. (*a*) 20 dB; (*b*) 220 K

12.13. (*a*) 8.45 dB; (*b*) 1740 K

12.15. 940.8 K

12.19. 108.1 dBHz; 32.5 dB

12.21. 6 dB

12.23. (*a*) 178.6 pW/m^2; (*b*) −97.5 dBW/m^2

12.25. 45.6 dBW

12.27. 78.1 dBHz

12.29. 73.5 dBW

12.31. 92.1 dBHz; 86.1 dBHz

12.33. 14.6 dB

12.35. 76.1 dBHz

12.37. 14.5 dB

12.39. 26.52 dB

Chapter 13

13.3. For earth station 1, 2.83°, and for earth station 2, 2.85°

13.5. 0.5°; 314.2 km

13.7. 4.5 dB

13.9. 45.4 dB

13.11. 25.1 dB

13.13. 21.1 dB

13.15. 51.35 dB

13.17. Yes for 13.12; no for 13.13

13.19. (*a*) 22.4 dB; (*b*) 25 dB

Chapter 14

14.9. 6

14.11. 26.5 dB

14.13. (*a*) −146.8 $dBWK^{-1}$; (*b*) 15.7 dBW

14.15. (*b*)(1) 2.2 MHz; (*b*)(2) 11.2 MHz

14.21. 0.85

14.23. 0.51

14.25. 0.93; 872

14.27. 0.962; 1797

14.33. 60 Mb/s

14.35. (*a*) 75.89 dBHz; (*b*) 49.3 dBW

14.37. 50.5 dBW

14.39. 24 modes

14.41. 8 mV; −21 dB

14.49. 199 (rounded down)

Chapter 15

15.11. 33.92 Mbps

15.23. 40 kb/s

15.25. 251.7 Mb/s

15.31. 1904 kilobytes

15.33. 0.327 s; 0.267 s

Chapter 16

16.1. (*a*) 22 at 148°, 10 unassigned; 30 at 166°, 2 unassigned; 27 at 157°, 5 unassigned; 32 at 119°, 110°, and 101°; 30 at 61.5°,

2 unassigned. (*b*) Continental 22; DBSC 22; DirecTV 54; Dominion 16; Echostar 24; Echostar/Directsat 44; MCI 28; TIC/Tempo 11; USSB 16; unassigned 19.

16.3. 54 dBW

16.5. **175°:** DBSC 440 Mb/s; Echostar/Directsat 440 Mb/s

166°: Continental 440 Mb/s; Dominion 320 Mb/s; Echostar/ Directsat 440 Mb/s

157°: DirecTV 1080 Mb/s

148°: Echostar 960 Mb/s; USSB 320 Mb/s

119°: Echostar/Directsat 840 Mb/s; TCI/Tempo 440 Mb/s

110°: Echostar/Directsat 40 Mb/s; MCI 1120 Mb/s; USSB 120 Mb/s

101°: DirecTV 1080 Mb/s; USSB 200 Mb/s

61.5°: Continental 440 Mb/s; DBSC 440 Mb/s; Dominion 320 Mb/s

16.13. 31 dB; 33 dB; 34 dB; 3.67°; 2.76°; 2.21°

16.15. $[E_b/N_0] = -0.46$ dB; therefore, signal is completely lost.

16.17. 6.9 dB. This is just satisfactory.

16.19. With an antenna diameter of 20 in., a rain rate of 18 mm/h will just lose the signal. This corresponds to a time percentage of about 0.49 percent (see Table 16.2).

16.21. 11.48 MHz; 33.44 MHz

Chapter 17

(Values rounded off to 1 decimal place).

17.7. 144.6 dB

17.9. 71.1 dBHz; 23.5 dB

17.11. 57.3 dBHz; 20.5 dB

17.13. 56.2 dBHz and 9.2 dB for both cells

17.15. Cell-1:51.4 dBHz, 4.4 dB; cell-16:52.1 dBHz, 5.1 dB

17.17. Rain conditions 4.8 dB; clear sky conditions 7.8 dB

17.19. No sun: 78.9 dBHz, 4.9 dB; sun: 78.5 dBHz, 4.6 dB.

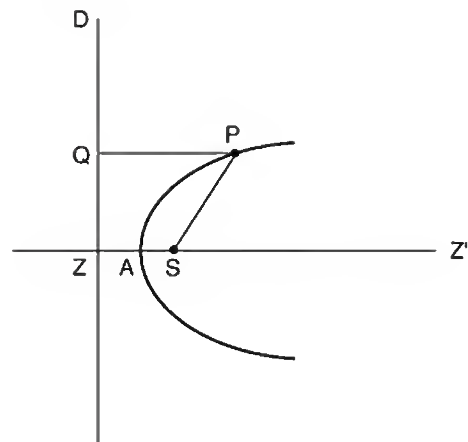
Conic Sections

A *conic section*, as the name suggests, is a section taken through a cone. At the intersection of the sectional plane and the surface of the cone, curves having many different shapes are produced, depending on the inclination of the plane, and it is these curves which are referred to generally as conic sections. Although the origin of conic sections lies in solid geometry, the properties are readily expressed in terms of plane geometrical curves. In Fig. B.1*a*, a reference line for conic sections, known as the *directrix*, is shown as *Z-D*. The *axis* for the conic sections is shown as line *Z-Z'*. The axis is perpendicular to the directrix. The point *S* on the axis is called the *focus*. For all conic sections, the focus has the particular property that the ratio of the distance *SP* to distance *PQ* is a constant. *SP* is the distance from the focus to any point *P* on the curve (conic section), and *PQ* is the distance, parallel to the axis, from point *P* to the directrix. The constant ratio is called the *eccentricity*, usually denoted by *e*. Referring to Fig. B.1*a*,

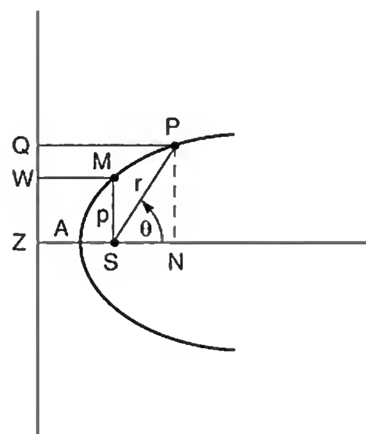
$$e = \frac{SP}{PQ} \tag{B.1}$$

The conic sections are given particular names according to the value of *e*, as shown in the following table:

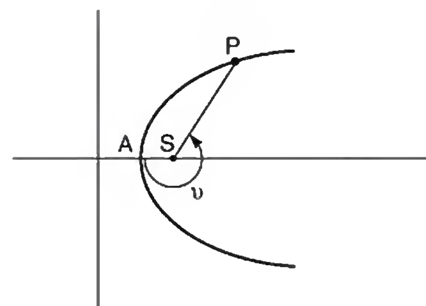
Curve	Eccentricity, *e*
Ellipse	$e < 1$
Parabola	$e = 1$
Hyperbola	$e > 1$



(a)



(b)



(c)

Figure B.1

These curves are encountered in a number of situations. In this book they are used to describe:

1. The path of satellites orbiting the earth
2. The ellipsoidal shape of the earth
3. The outline curves for various antenna reflectors

A polar equation for conic sections with the pole at the foci can be obtained in terms of the fixed distance p, called the *semilatus rectum*. The polar equation relates the point P to the radius r and the angle θ Fig. B.1*b*). From Fig. B.1*b*,

$$\begin{aligned} SN &= ZN - ZS \\ &= QP - WM \\ &= \frac{r}{e} - \frac{p}{e} \end{aligned} \tag{B.2}$$

Also,

$$SN = r\cos\theta \tag{B.3}$$

Combining Eqs. (B.2) and (B.3) and simplifying gives the polar equation as

$$r = \frac{p}{1 - e\cos\theta} \tag{B.4}$$

If the angle is measured from SA, shown as ν in Fig. B.1*c*, then, since $\nu = 180° + \theta$, the polar equation becomes

$$r = \frac{p}{1 + e\cos\nu} \tag{B.5}$$

The Ellipse

For the ellipse, $e < 1$. Referring to Eq. (B.4), when $\theta = 0°$, $r = p/(1 - e)$. Since $e < 1$, r is positive. At $\theta = 90°$, $r = p$, and at $\theta = 180°$, $r = p/(1 + e)$. Thus r decreases from a maximum of $p/(1 - e)$ to a minimum of $r = p/(1 + e)$, the locus of r describing the closed curve $A'BMA$ (Fig. B.2). Also, since $\cos(-\theta) = \cos\theta$, the curve is symmetrical about the axis, and the closed figure results (Fig. B.2*a*).

The length AA' is known as the *major axis* of the ellipse. The semimajor axis $a = AA'/2$ and e are the parameters normally specified for an ellipse. The semilatus rectum p can be obtained in terms of these two quantities. As already shown, the maximum value for r is $SA' = p/(1 - e)$ and the

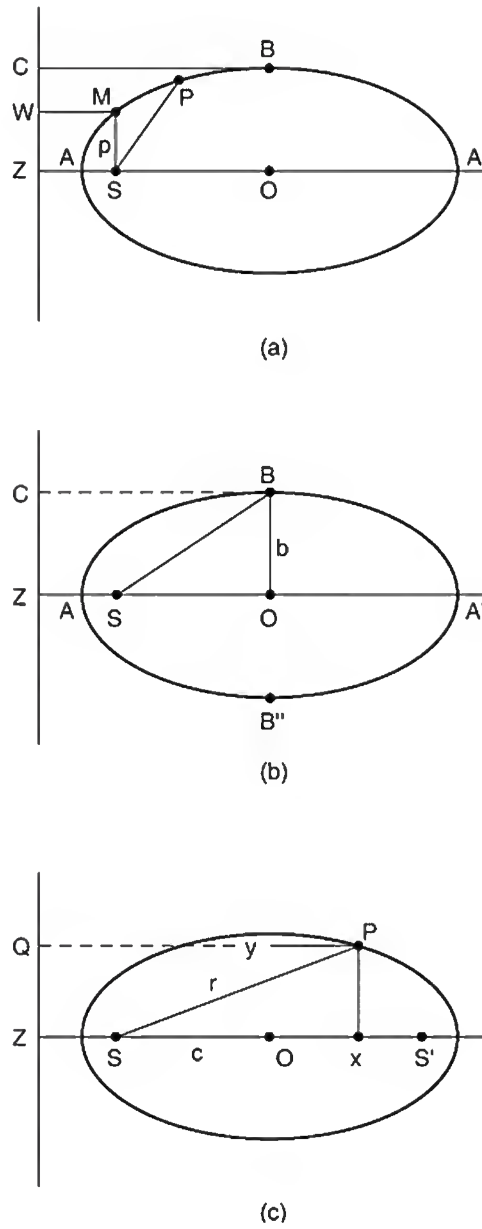


Figure B.2

minimum value is $SA = p/(1 + e)$. Adding these two values gives

$$\begin{aligned} AA' &= AS + SA' \\ &= \frac{p}{1+e} + \frac{p}{1-e} \\ &= \frac{2p}{1-e^2} \end{aligned}$$

Since $a = AA'/2$, it follows that

$$p = a(1 - e^2) \tag{B.6}$$

Substituting this into Eq. (B.5) gives

$$r = \frac{a(1 - e^2)}{1 + e\cos\nu} \tag{B.7}$$

which is Eq. (2.23) of the text.

Equation (B.7) can be written as

$$r = \frac{a(1 + e)(1 - e)}{1 + e\cos\nu}$$

When $\nu = 360°$,

$$r = a(1 - e) \tag{B.8}$$

which is Eq. (2.6) of the text. When $\nu = 180°$

$$r = a(1 + e) \tag{B.9}$$

which is Eq. (2.5) of the text.

Referring again to Fig. B.2*a*, and denoting the length $SO = c$, it is seen that $AS + c = a$. But, as shown above, $AS = p/(1 + e)$ and $p = a(1 - e^2)$. Substituting these for AS and simplifying gives

$$c = ae \tag{B.10}$$

Point O bisects the major axis, and length BO is called the *semiminor axis*, denoted by b (BB'' is the minor axis) (Fig. B.2*b*).

The semiminor axis can be found in terms of a and e as follows: Referring to Fig. B.2*b*, $2a = SA + SA' = e(AZ + A'Z) = e(2OZ)$, and therefore,

$$OZ = \frac{a}{e} \tag{B.11}$$

But $OZ = BC = SB/e$, and therefore, $SB = a$. SB is seen to be the radius at B. From the right-angled triangle so formed,

$$\begin{aligned} a^2 &= b^2 + c^2 \\ &= b^2 + (ae)^2 \end{aligned}$$

From this it follows that

$$b = a\sqrt{1 - e^2} \tag{B.12}$$

and

$$e = \frac{\sqrt{a^2 - b^2}}{a} \tag{B.13}$$

This is Eq. (2.1) of the text.

The equation for an ellipse in rectangular coordinates with origin at the center of the ellipse can be found as follows: Referring to Fig. B.2*c*, in which *O* is at the zero origin of the coordinate system,

$$r^2 = (c + x)^2 + y^2$$

But $r = ePQ = e(OZ + x) = e(a/e + x) = a + ex$. Hence

$$(a + ex)^2 = (c + x)^2 + y^2$$

Multiplying this and simplifying gives

$$a^2(1 - e^2) = x^2(1 - e^2) + y^2$$

Hence,

$$\begin{aligned} 1 &= \frac{x^2}{a^2} + \frac{y^2}{a^2(1 - e^2)} \\ &= \frac{x^2}{a^2} + \frac{y^2}{b^2} \end{aligned} \tag{B.14}$$

This shows the symmetry of the ellipse, since for a fixed value of y the x^2 term is the same for positive and negative values of x. Also, because of the symmetry, there exists a second directrix and focal point to the right of the ellipse. The second focal point is shown as S' in Fig. B.2*c*. This is positioned at $x = c = ae$ (from Eq. B.10).

In the work to follow, y can be expressed in terms of x as

$$y = \frac{b}{a}\sqrt{a^2 - x^2} \tag{B.15}$$

To find the area of an ellipse, consider first the area of any segment (Fig. B.3). The area of the strip of width dx is $dA = y\,dx$, and hence the area ranging from $x = 0$ to x is

$$\begin{aligned} A_x &= \int_0^x y\mathrm{dx} \\ &= \int_0^x \frac{b}{a}\sqrt{a^2 - x^2}dx \end{aligned} \tag{B.16}$$

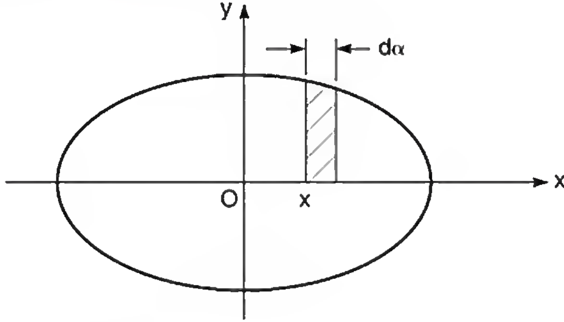


Figure B.3

This is a standard integral which has the solution

$$A_x = \frac{xy}{2} + \frac{ab}{2}\phi \tag{B.17}$$

where $\phi = \arcsin(x/a)$. In particular, when $x = a$, $\phi = \pi/2$, and the area of the quadrant is

$$A_q = \frac{ab\pi}{4} \tag{B.18}$$

It follows that the total area of the ellipse is

$$A = 4A_q = \pi ab \tag{B.19}$$

In satellite orbital calculations, time is often measured from the instant of perigee passage. Denote the time of perigee passage as T and any instant of time after perigee passage as t. Then the time interval of significance is $t - T$. Let A be the area swept out in this time interval, and let T_p be the periodic time. Then, from Kepler's second law,

$$A = \pi ab\frac{t - T}{T_p} \tag{B.20}$$

The mean motion is $n = 2\pi/T_p$ and the mean anomaly is $M = n\,(t - T)$. Combining these with Eq. (B.20) gives

$$A = \frac{Mab}{2} \tag{B.21}$$

The auxiliary circle is the circle of radius a circumscribing the ellipse as shown in Fig. B.4. This also shows the *eccentric anomaly*, which is angle E, and the true anomaly v (both of which are measured from perigee). The true anomaly is found through the eccentric anomaly.

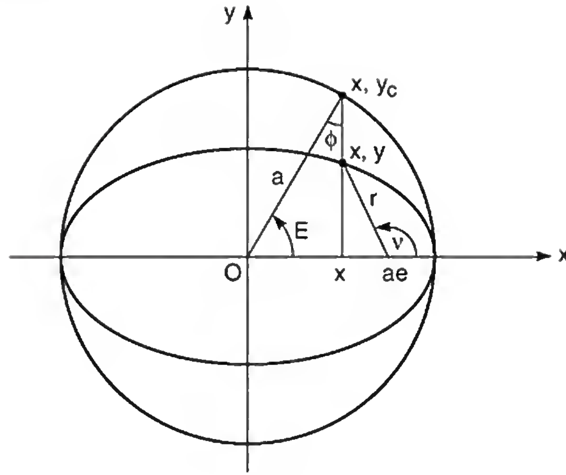


Figure B.4

Some relationships of importance which can be seen from the figure are

$$E = \frac{\pi}{2} - \phi \tag{B.22}$$

$$y_c = a \sin E \tag{B.23}$$

The equation for the auxiliary circle is $x^2 + y_c^2 = a^2$. Substituting for x^2 from this into Eq. (B.15) and simplifying gives

$$y = \frac{b}{a} y_c \tag{B.24}$$

Combining this with Eq. (B.23) gives another important relationship:

$$y = b \sin E \tag{B.25}$$

The area swept out in time $t - T$ can now be found in terms of the individual areas evaluated. Referring to Fig. B.5, this is

$$\begin{aligned} A &= A_q - A_x - A_\Delta \\ &= \frac{\pi ab}{4} - \left(\frac{xy}{2} + \frac{ab}{2}\phi\right) - \frac{(ae - x)}{2} y \\ &= \frac{ab}{2}\left(\frac{\pi}{2} - \phi - e \sin E\right) \\ &= \frac{ab}{2}(E - e \sin E) \end{aligned} \tag{B.26}$$

Comparing this with Eq. (B.21) shows that

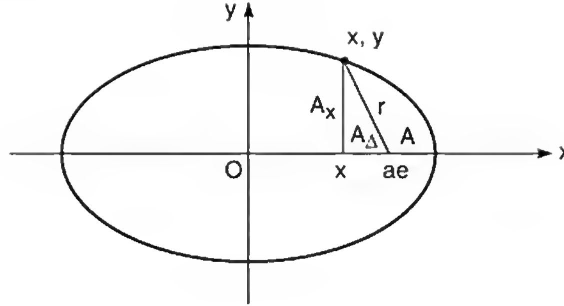


Figure B.5

$$M = E - e\sin E \tag{B.27}$$

This is Kepler's equation, given as Eq. (2.27) in the text.

The orbital radius r and the true anomaly v can be found from the eccentric anomaly. Referring to Fig. B.4,

$$r\cos(180° - \nu) = ae - x$$

But $x = a \cos E$, and hence,

$$r\cos\nu = a(\cos E - e) \tag{B.28}$$

Also, from Fig. B.4,

$$r\sin(180° - \nu) = y$$

and as previously shown, $y = b \sin E$ and $b = a\sqrt{1 - e^2}$. Hence,

$$r\sin\nu = a\sqrt{1 - e^2}\sin E \tag{B.29}$$

Squaring and adding Eqs. (B.29) and (B.28) gives

$$r^2 = a^2(\cos E - e)^2 + a^2(1 - e^2)\sin^2 E$$

from which

$$r = a(1 - e\cos E) \tag{B.30}$$

This is Eq. (2.30) of the text.

One further piece of manipulation yields a useful result. Combining Eqs. (B.28) and (B.30) gives

$$\cos\nu = \frac{\cos E - e}{1 - e\cos E}$$

and hence

$$\frac{1 - \cos\nu}{1 + \cos\nu} = \frac{1 - \dfrac{\cos E - e}{1 - e\cos E}}{1 + \dfrac{\cos E - e}{1 - e\cos E}}$$

$$= \frac{(1 + e)(1 - \cos E)}{(1 - e)(1 + \cos E)}$$

Using the trigonometric identity for any angle α that

$$\tan^2 \frac{\alpha}{2} = \frac{1 - \cos\alpha}{1 + \cos\alpha}$$

yields

$$\tan\frac{\nu}{2} = \sqrt{\frac{1 + e}{1 - e}}\tan\frac{E}{2} \qquad \text{(B.31)}$$

This is Eq. (2.29) of the text.

An elliptical reflector has focusing properties, which may be derived as follows. From Fig. B.6,

$$OZ = a + AZ$$

$$= a + \frac{AS}{e}$$

But $AS = a(1 - e)$, as shown by Eq. (B.8); hence,

$$OZ = \frac{a}{e}$$

Referring again to Fig. B.6, $SP = eQP$, and $S'P = ePQ'$. Hence

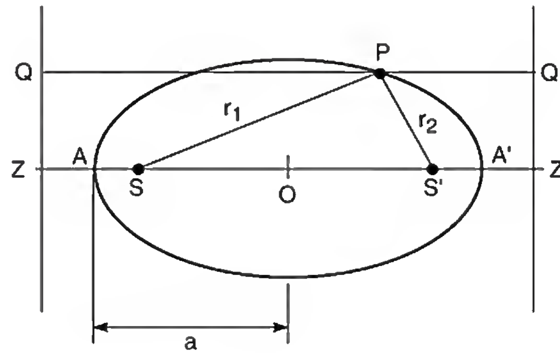


Figure B.6

$$\begin{aligned} SP + S'P &= e(QP + PQ') \\ &= e(ZZ') \\ &= e(2 \times OZ) \\ &= 2a \end{aligned}$$

But $SP = r_1$ and $S'P = r_2$, and hence,

$$r_1 + r_2 = 2a \tag{B.32}$$

This shows that the sum of the focal distances is constant, or in other words, the ray paths from one focus to the other which go via an elliptical reflector are equal in length. Thus, electromagnetic radiation emanating from a source placed at one of the foci will have the same propagation time over any reflected path, and therefore, the reflected waves from all parts of the reflector will arrive at the other foci in phase. This property is made use of in the Gregorian reflector antenna described in Sec. 6.15.

The Parabola

For the parabola, the eccentricity $e = 1$. Referring to Fig. B.7, since $e = SA/AZ$, by definition it follows that $SA = AZ$. Let $f = SA = AZ$. This is known as the *focal length*. Consider a line LP' drawn parallel to the directrix. The path length from S to P' is $r + PP'$. But $PP' = AL - f - r \cos\theta$, and hence the path length is $r(1 - \cos\theta) + AL - f$. Substituting for r from Eq. (B.4) with $e = 1$ yields a path length of $p + AL - f$. This shows that the path length of a ray originating from the focus and reflected parallel

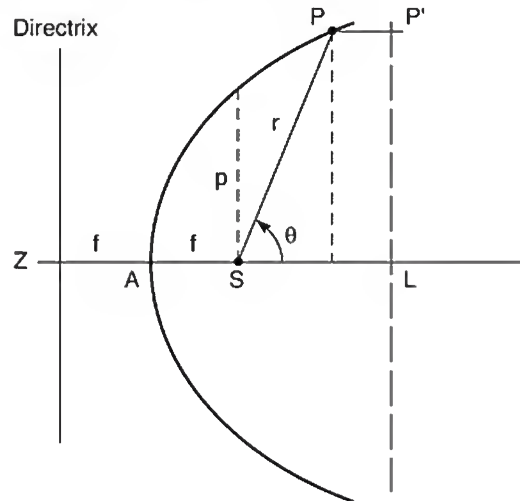


Figure B.7

to the axis (*ZZ′*) is constant. It is this property which results in a parallel beam being radiated when a source is placed at the focus.

It will be seen that when $\theta = 90°$, $r = SP = p$. But $SP = eSZ = 2f$ (since $e = 1$), and therefore, $p = 2f$. Thus, the radius as given by Eq. (B.4) can be written as

$$r = \frac{2f}{1 - \cos\theta} \tag{B.33}$$

This is essentially the same as Eq. (6.27) of the text, where ρ is used to replace r, and $\Psi = 180° - \theta$, and use is made of the trigonometric identity $1 + \cos\Psi = 2\cos^2(\Psi/2)$. Thus

$$\frac{\rho}{f} = \sec^2\frac{\Psi}{2} \tag{B.34}$$

For the situation shown in Fig. B.8, where D is the diameter of a parabolic reflector, Eq. (B.34) gives

$$\frac{f}{\rho_0} = \cos^2\frac{\Psi_0}{2} \tag{B.35}$$

But from Fig. B.8 it is also seen that

$$\rho_0 = \frac{D}{2\sin\Psi_0} = \frac{D}{4\sin\dfrac{\Psi_0}{2}\cos\dfrac{\Psi_0}{2}}$$

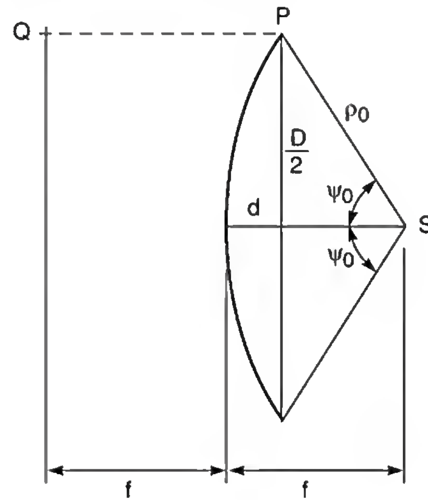


Figure B.8

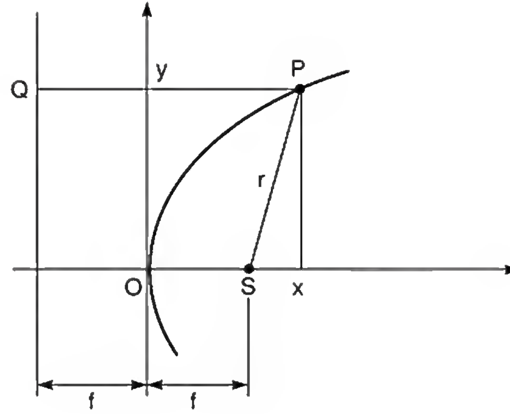


Figure B.9

where use is made of the double angle formula sin $\Psi_0 = 2\sin(\Psi_0/2)\cos(\Psi_0/2)$. Substituting for ρ_0 in Eq. (B.34) and simplifying gives

$$\frac{f}{D} = \frac{1}{4}\cot\frac{\Psi_0}{2} \tag{B.36}$$

This is Eq. (6.29) of the text.

It is sometimes useful to be able to locate the focal point, knowing the diameter D and the depth d of the dish. From the property of the parabola, $\rho_0 = PQ = f + d$. From Fig. B.8 it is also seen that $\rho_0^2 = (D/2)^2 + (f - d)^2$. Substituting for ρ_0 and simplifying yields

$$f = \frac{D^2}{16d} \tag{B.37}$$

With the zero origin of the xy coordinate system at the vertex (point A) of the parabola, the line PQ becomes equal to $x + f$ (Fig. B.9), and $y = r^2 - (x - f)^2$. These two results can be combined to give the equation for the parabola:

$$y^2 = 4fx \tag{B.38}$$

The Hyperbola

For the hyperbola, $e > 1$. The curve is sketched in Fig. B.10. Equation (B.4) still applies, but it will be seen that for $\cos\theta = 1/e$ the radius goes to infinity; in other words, the hyperbolic curve does not close on itself in the way that the ellipse does but lies parallel to the radius at this value of θ. In constructing the ellipse, because e was less than unity, it was possible to find a point A' to the right of S for which the ratio $SA'/A'Z = e$ applied in addition to the ratio $e = SA/AZ$. With the hyperbola, because $e > 1$, a point A'

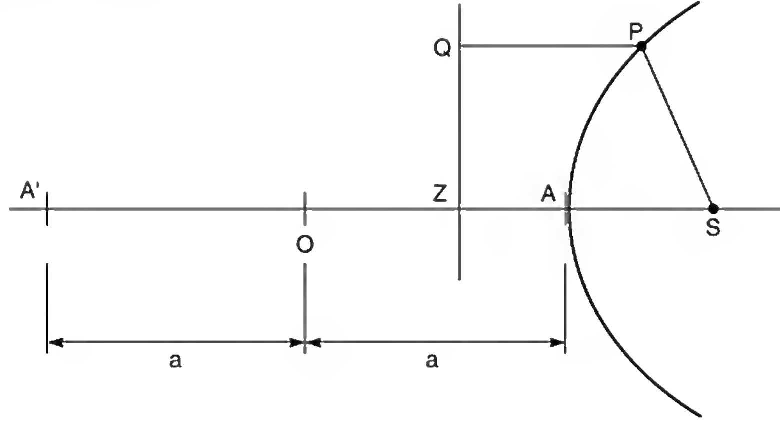


Figure B.10

to the left of S can be found for which $e = SA'/A'Z$ in addition to $e = SA/AZ$. By setting $2a$ equal to the distance $A'A$ and making the midpoint of $A'A$ equal to the x, y coordinate origin O as shown in Fig. B.10, it is seen that $2OS = SA + SA'$. But $SA' = eA'Z$ and $SA = eZA$. Hence,

$$2OS = e(ZA + A'Z)$$
$$= 2ae$$

Hence,

$$OS = ae \tag{B.39}$$

Also, $SA' - SA = 2a$, and hence

$$2a = e(A'Z - ZA)$$
$$= e2OZ$$

Therefore,

$$OZ = \frac{a}{e} \tag{B.40}$$

Point P on the curve can now be given in terms of the xy coordinates. Referring to Fig. B.11, S is at point ae, and

$$SP^2 = (ae - x)^2 + y^2$$

Also,

$$SP = ePQ = e\left(x - \frac{a}{e}\right)$$

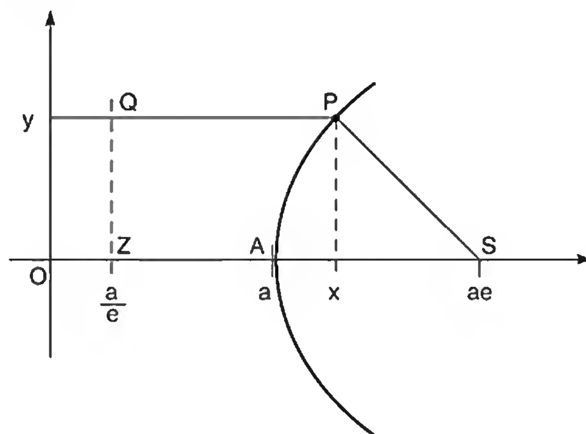


Figure B.11

Combining these two equations and simplifying yields

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \tag{B.41}$$

where in this case

$$b^2 = a^2 (e^2 - 1) \tag{B.42}$$

By plotting values it will be seen that a symmetrical curve results, as shown in Fig. B.12, and that there is a second focus at point *S'*. An important property of the hyperbola is that the difference of the two focal distances is a constant. Referring to Fig. B.12, $S'P = ePQ'$ and $SP = ePQ$.

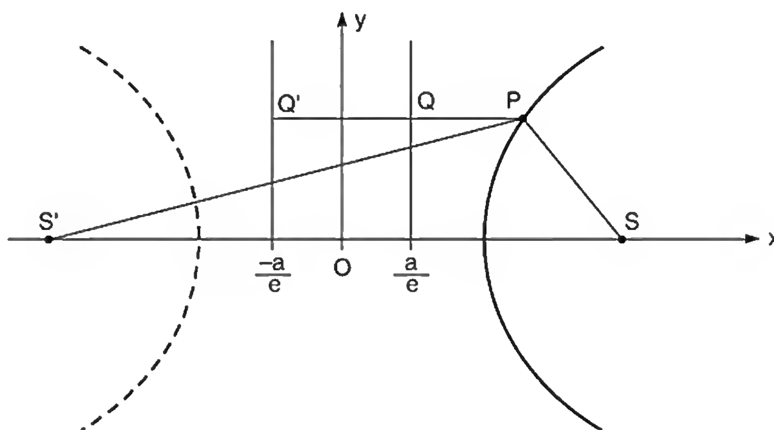


Figure B.12

Hence,

$$S'P - SP = e(PQ' - PQ) = e\left(\frac{2a}{e}\right) = 2a \tag{B.43}$$

An application of this property is shown in Fig. 6.23*a* of the text which is redrawn in Fig. B.13. By placing the focus of the parabolic reflector at the focus *S* of the hyperbola and the primary source at the focus *S*′, the total path length $S'P + PP''$ is equal to $2a + SP + PP''$ or $2a + SP''$. But, as shown previously in Fig. B.7, the focusing properties of the parabola rely on there being a constant path length $SP'' + P''P'$, and adding the constant $2a$ to SP'' does not destroy this property.

The double-reflector arrangement can be analyzed in terms of an *equivalent parabola*. The equivalent parabola has the same diameter as the real parabola and is formed by the locus of points obtained at *P*′ which is the intersection of $S'P$ produced to *P*′ and $P''P'$ which is parallel to the *x* axis, as shown in Fig. B.14. The focal distance of the equivalent parabola is shown as f_e and of the real parabola as *f*. Looking from the focus *S* to the real parabola, one sees that

$$\frac{h}{d_2} = \frac{y}{f - X_1}$$

Looking to the right from focus *S*′ to the equivalent parabola, one sees that

$$\frac{h}{d_1} = \frac{y}{f_e - X_2}$$

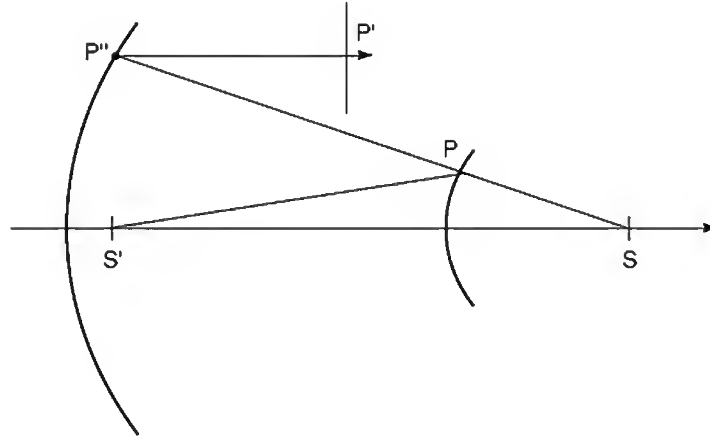


Figure B.13

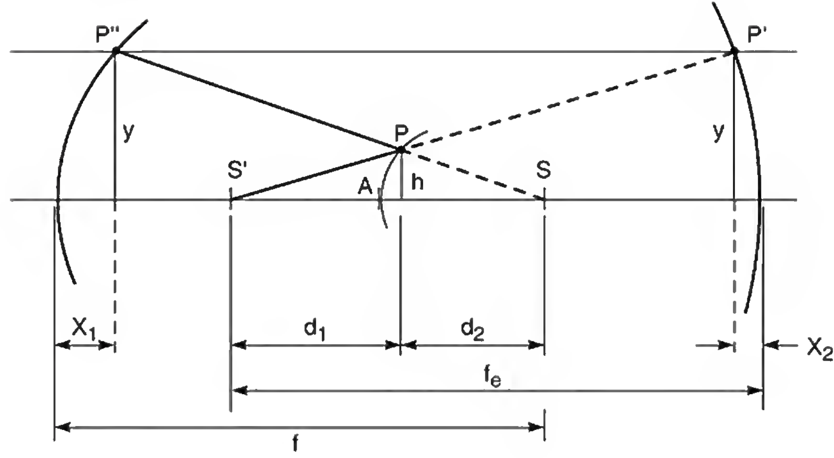


Figure B.14

Hence, equating h/y from these two results gives

$$\frac{d_1}{f_e - X_2} = \frac{d_2}{f - X_1}$$

In the limit as h goes to zero, X_1 and X_2 both go to zero, and d_1 goes to $S'A$ and d_2 to AS. Hence,

$$\frac{S'A}{f_e} = \frac{AS}{f}$$

From Fig. B.12 and Eq. (B. 39),

$$SS' = 2OS = 2ae$$

From Fig. B.11,

$$AS = ae - a = a(e - 1)$$

From Fig. B.14,

$$S'A = SS' - AS = a(e + 1)$$

Hence,

$$\frac{a(e + 1)}{f_e} = \frac{a(e - 1)}{f}$$

From which

$$f_e = \frac{e + 1}{e - 1} f \qquad \text{(B.44)}$$

This is Eq. (6.35) of the text.

NASA Two-Line Orbital Elements

The two-line orbital elements can be found at a number of Web sites. The *National Oceanographic and Atmospheric Administration* (NOAA) Web site, at http://www.noaa.gov/, is probably the most useful to start with because it contains a great deal of general information on polar orbiting satellites as well as weather satellites in the geostationary orbit. An explanation of the two-line elements can be found in the FAQs section by Dr. T. S. Kelso at http://celestrak.com/. The two-line elements can be downloaded directly from http://celestrak.com/NORAD/elements/, (but see App. D) a typical readout being:

1 22969U 94003A 94284.57233250 .00000051 00000-0 10000-3 0 1147
2 22969 82.5601 334.1434 0015195 339.6133 20.4393 13.16724605 34163

A description of each line follows:

Line Number 1

Name	Description	Units	Example	Field Format
LINNO	Line number of element data (always 1 for line 1)	None	1	X
SATNO	Satellite number	None	22969	XXXXX
U	Not applicable	None		X
IDYR	International designator (last two digits of launch year)	Launch year	94	XX
IDLNO	International designator (launch number of the year)	None	3	XXX
EPYR	Epoch year (last two digits of the year)	Epoch year	94	XX
EPOCH	Epoch (Julian day and fractional portion of the day)	Day	284.57233250	XXX.XXXXXXXX
NDTO2 or BTERM	First time derivative of the mean motion or ballistic coefficient (depending on the ephemeris type)	Revolutions per day^2 or m^2/kg	0.00000051	±.XXXXXXXX*
NDDOT 6	Second time derivative of mean motion (field will be blank if NDDOT6 is not applicable)	Revolutions per day^3	00000-0	±XXXXX-X†
BSTAR or AGOM.	BSTAR drag term if GP4 general perturbations theory was used. Otherwise it will be the radiation pressure coefficient		10000-3	±XXXXX-X

Line Number 1 (*Continued*)

EPHTYP	Ephemeris type (specifies the ephemeris theory used to produce the elements)	None	0	X
ELNO	Element number	None	1147	XXXX

*If NDOT2 is greater than unity, a positive value is assumed without a sign.
†Decimal point assumed after the ± signs.

Line Number 2

Name	Description	Units	Example	Field Format
LINNO	Line number of element data (always 2 for line 2)	None	2	X
SATNO	Satellite number	None	22969	XXXXX
II	Inclination	Degrees	82.5601	XXX.XXXX
NODE	Right ascension of the ascending node	Degrees	334.1434	XXX.XXXX
EE	Eccentricity (decimal point assumed)	None	00151950	XXXXXXXX
OMEGA	Argument of perigee	Degrees	339.6133	XXX.XXXX
MM	Mean anomaly	Degrees	20.4393	XXX.XXXX
NN	Mean motion	Revolutions per day	13.16724605	XX.XXXXXXXX
REVNO	Revolution number at epoch	Revolutions	34163	XXXXX

Listings of Artificial Satellites

The best source by far has been that provided by Dr. T. S. Kelso at Celestrak, at http://celestrak.com/. The source of CelesTrak's data was the NASA *orbital information group* (OIG), but it appears that this site will be closed when the *Air Force Space Command* (AFSC) has in operation its program to provide space surveillance data—including NORAD *two-line element sets* (TLEs)—to *non-US government entities* (NUGE). A new site is available at http://www.space-track.org/, but only approved registered users can access this site. Registration is quite straightforward.

Other useful sites are:

http://www.amsat.org/amsat/keps/formats.html
http://web.austin.utexas.edu/
http://www.lyngsat.com/

In general, a web search by satellite name or by entering a search for two-line elements, for example, will provide a list of useful web sites.

Illustrating Third-Order Intermodulation Products

The nonlinear voltage transfer characteristic for a TWT can be written as a power series

$$e_o = ae_i + be_i^2 + ce_i^3 + \cdots$$

The third-order term gives rise to the intermodulation products. To illustrate this, let the input be two unmodulated carriers

$$e_i = A\cos\omega_A t + B\cos\omega_B t$$

and let the carrier spacing be $\Delta\omega$ as shown in Fig. E.1. The terms on the right-hand side of the transfer characteristic equation can be interpreted as follows:

First term, ae_i. This gives the desired linear relationship between e_o and e_i.

Second term, be_i^2. With e_i as shown, this term can be expanded into the following components: a dc component, a component at frequency $\Delta\omega$, second harmonic components of the carriers, second harmonic + $\Delta\omega$ components. All these components can be removed by filtering and need not be considered further.

Third term, ce_i^3. This can be expanded as

$$c(A^3\cos^3\omega_A t + B^3\cos^3\omega_B t + 3A^2B\cos^2\omega_A t\cos\omega_B t + 3AB^2\cos\omega_A t\cos^2\omega_B t)$$

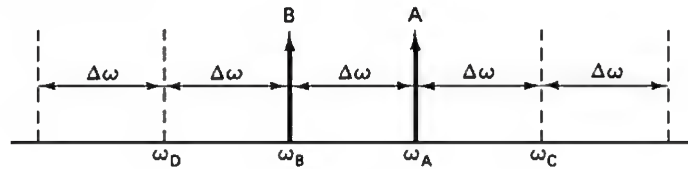


Figure E.1

The cubed terms in this can be expanded as

$$\cos^3 \omega t = \frac{1}{4}(3\cos\omega t + \cos 3\omega t)$$

that is, a fundamental component plus a third harmonic component. The third harmonics can be removed by filtering.

The intermodulation products are contained in the cross-product terms $3A^2B\ \cos^2\omega_A t\ \cos\omega_B t$ and $3AB^2\ \cos\omega_A t\ \cos^2\omega_B t$. On further expansion, the first of these will be seen to contain a $\cos(2\omega_A - \omega_B)t$. With the carriers spaced equally by amount $\Delta\omega$, the $2\omega_A - \omega_B$ frequency is equal to $\omega_A + (\omega_A - \omega_B) = \omega_A + \Delta\omega$. This falls exactly on the adjacent carrier frequency at ω_C, as shown in Fig. E.1. Likewise, the expansion of the second cross-product term contains a $\cos(2\omega_B - \omega_A)t$ which yields an intermodulation product at frequency $\omega_B - \Delta\omega$. This falls exactly on the carrier frequency at ω_D.

Acronyms

AAL	ATM Adaptation layer
ABR	Available bit rate
ACK	Acknowledgment
ACSSB	Amplitude companded single-sideband
AFSC	Air force space command
ALA	ATM link accelerator
AM	Amplitude modulation
AMPS	Analog mobile phone service
AMSC	American Mobile Satellite Consortium
ANSI	American national standards institute
AOS	Acquisition of signal
APM	Amplitude phase modulation
ATM	Asynchronous transfer mode; Asynchronous transmission mode
ATSC	Advanced television systems committee
AVC	Advanced video coding
BAPTA	Bearing and power transfer assembly
BDP	Bandwidth delay product
BER	Bit error rate
BOL	Beginning of life
BPSK	Binary phase-shift keying
BSS	Broadcasting satellite service
BSU	Burst synchronization unit
BT	Burst tolerance

CAN	Campus area network
CBR	Constant bit rate
CCIR	Comite Consultatif International de Radio
CCITT	Comite Consultatif International de Telephone et Telegraph
CDV	Cell delay variation
CER	Cell error ratio
CFM	Companded frequency modulation
CLP	Cell loss priority
CLR	Cell loss ratio
COMSAT	Communications Satellite (Corporation)
CONUS	Contiguous United States
CPFSK	Continuous phase frequency shift keying
CS	Convergence sublayer
CSSB	Companded single sideband
CTD	Cell transfer delay
DAMA	Demand assignment multiple access
DBS	Direct broadcast satellite
DCS	Digital cross-connect switch
DCT	Discrete cosine transform
DM	Delta modulation
DOMSAT	Domestic satellite
DPCM	Differential PCM
DSI	Digital speech interpolation
DSS	Digital satellite services
DTV	Digital television
DVB	Digital video broadcast
DVD	Digital versatile disk
ECS	European communications satellite
EDTV	Enhanced definition television
EHF	Extremely high frequency
EIRP	Equivalent isotropically radiated power
EOL	End of life
ERS	Earth resources satellite
ESA	European Space Agency
ETSI	European telecommunications standards institute
EUMETSAT	European organization for the exploration of METSATs

FCC	Federal Communications Commission
FDM	Frequency-division multiplex
FDMA	Frequency-division multiple access
FEC	Forward error correction
FFSK	Fast FSK
FM	Frequency modulation
FSK	Frequency-shift keying
FSL	Free-space loss
FSS	Fixed satellite service
GEO	Geostationary earth orbit
GFC	Generic flow control
GHz	Gigahertz
GOES	Geostationary operational environmental satellite
GPS	Global Positioning Satellite
GSM	Global system for mobile communications (originally Groupe Spe'cial Mobile)
GSO	Geo-synchronous orbit
HAN	Home area network
HDTV	High-definition television
HEC	Header error control
HEO	Highly elliptical orbit
HP	Horizontal polarization
HPA	High-power amplifier
HPBW	Half-power beamwidth
IDU	Indoor unit
IEC	International Electrotechnical Commission
IEE	Institution of Electrical Engineers
IEEE	Institute of Electrical and Electronic Engineers
IMMARSAT	International Maritime Satellite (Organization)
INTELSAT	International Telecommunications Satellite (Organization)
INTERSPUTNIK	International Sputnik (Satellite Communications)
IOT	In-orbit testing
IP	Internet Protocol
IRD	Integrated receiver-decoder
ISBN	Integrated Satellite Business Network

ISL	Intersatellite link
ISO	International standards organization
ISP	Internet service provider
ISU	Iridium subscriber unit
ITU	International Telecommunications Union
ITU-T	ITU-telecommunication standardization sector
IWU	Interworking unit
JPEG	Joint Photographic Experts Group
LAN	Local area network
LANDSAT	Land survey satellite
LDPC	Low density parity check
LEO	Low earth orbit
LHCP	Left-hand circular polarization
LLR	Log likelihood ratio
LNA	Low-noise amplifier
LNB	Low-noise block
LOS	Loss of signal
LSAT	Large satellite (European)
MA	Multiple Access
MAC	Monitoring alarm and control
MAC	Multiplexed analog compression
MAN	Metropolitan area network
MARISAT	Marine satellite (communications)
MEO	Medium earth orbit
METOP	Meteorological operations
METSAT	Meteorological satellite
MHz	Megahertz
MPEG	Motion Pictures Expert Group
MSAT	Mobile satellite (for mobile communications)
MSK	Minimum shift keying
MUX	Multiplexer
NACK	Negative acknowledgment
NASA	National Aeronautics and Space Administration
NCC	Network control center

NCS	Network control station
NESS	National Earth Satellite Service
NNI	Network nod interface (also network network interface)
NOAA	National Oceanic and Atmospheric Administration
NORAD	North American Aerospace Defense Command
NPOESS	National polar operational environmental satellite system
NRT	Non real time
NSP	Network service provider
NTSC	National Television System Committee
OBP	On-board processing
ODU	Outdoor unit
OIG	Orbital information group
PAL	Phase altemation line
PC	Personal computer; personal communications
PCM	Pulse-code modulation
PCR	Peak cell rate
PLL	Phase-lock loop
PNNI	Private NNI
POES	Polar operational environmental satellite
POTS	Plain old telephone service
PSK	Phase-shift keying
PSTN	Public service telephone network
PTI	Payload type indicator
PVC	Permanent virtual circuit
PVP	Permanent virtual path
QoS	Quality of service
QPSK	Quaternary phase-shift keying
RADARSAT	Radar satellite (for remote sensing of earth's surface and atmosphere)
RDSS	Radio determination satellite service
RFC	Request for comments
RHCP	Right-hand circular polarization
RT	Real time

RTT	Round-trip time
SACK	Selective acknowledgment
SAR	Synthetic aperture radar
SARSAT	Search and rescue satellite
SATCOM	Satellite communications
SATM	Satellite ATM
SBS	Satellite Business Systems
SCPC	Single channel per carrier
SCR	Sustained cell rate
SDTV	Standard definition television
SECAM	Sequential couleur à memoire
SFD	Saturaton flux density
SHF	Super high frequencies
SI	Systeme International (d' Unites) (International System of Units)
SPADE	SCPC PCM multiple-access demand-assignment equipment
SPEC	Speech predictive encoded communications
SRB	Synchronization reference burst
SS	Satellite switched
SS	Spread spectrum
SSB	Single sideband
STM	Synchronous transmission mode
SV	Space vehicle
SVC	Switched virtual circuit
TASI	Time assignment speech interpolation
TCP	Transmission Control Protocol; Transport Control Protocol
TDM	Time-division multiplex
TDMA	Time-division multiple access
TDRSS	Tracking Data Radio Satellite System
TELESAT	Telecommunications satellite
TIROS	Television and infrared observational satellite
TLE	Two line elements
TMUX	Trans-MUX
TT&C	Tracking telemetry and command
TV	Television
TVRO	TV receive only

TWT	Traveling-wave tube
TWTA	Traveling-wave tube amplifier
UBR	Unspecified bit rate
UDP	User Datagram Protocol
UHF	Ultra high frequencies
UNI	User network interface
VBR	Variable bit rate
VCEG	Video coding experts group
VCI	Virtual channel identifier
VHF	Very high frequencies
VoIP	Voice over Internet Protocol
VP	Vertical polarization
VPI	Virtual path identifier
VSAT	Very small aperture terminal
WAN	Wide area network
WARC	World Administrative Radio Conference
WWW	World Wide Web
XPD	Cross-polarization discrimination

Logarithmic Units

Decibels

A power ratio of P_1/P_2 expressed in bels is

$$\log_{10}\left(\frac{P_1}{P_2}\right)$$

The bel was introduced as a logarithmic unit which allowed addition and subtraction to replace multiplication and division. As such, it proved to be inconveniently large, and the decibel, abbreviated dB, is the unit now in widespread use. The same power ratio expressed in decibels is

$$10\log_{10}\left(\frac{P_1}{P_2}\right)$$

Because the base 10 is understood, it is seldom, if ever, shown explicitly.

The abbreviation dB, as shown below, is often modified to indicate a reference level. Although referred to as a *unit*, the decibel is a dimensionless quantity, the *ratio* of two powers. Power by itself cannot be expressed in decibels. However, by selecting unit power as the denominator in the ratio, power can be expressed in decibels relative to this. A power expressed in decibels relative to 1 W is shown as dBW. For example, 50 W expressed in decibels relative to 1 W would be equivalent to

$$10\log\frac{50\text{W}}{1\text{W}} \cong 17\ \text{dBW}.$$

Another commonly used reference is the milliwatt, and decibels relative to this are shown as dBm. Thus 50 W relative to 1 mW would be

$$10\log\frac{50}{10^{-3}} \cong 47 \text{ dBm}$$

By definition, the ratio of two voltages V_1 and V_2 expressed in decibels is

$$20\log\frac{V_1}{V_2} \text{ dB}$$

The multiplying factor of 20 comes about because power is proportional to voltage squared. Note carefully, however, that the definition of voltage decibels stands on its own and can be related to the corresponding power ratio only when the ratio of load resistances is also known. Also, unit voltage may be selected as a reference, which allows voltage to be expressed in decibels relative to it. For example, 0.5 V expressed in decibels relative to 1 V is

$$20\log\frac{0.5 \text{ V}}{1 \text{ V}} \cong -6 \text{ dBV}$$

Another common voltage reference is the microvolt, and 0.5 V expressed in decibels relative to 1 μV is

$$20\log\frac{0.5 \text{ V}}{10^{-6} \text{ V}} \cong 114 \text{ dB}\mu\text{V}$$

Current ratios may also be expressed in decibels as:

$$20\log\frac{I_1}{I_2}$$

And a reference current may be chosen for I_2.

Decilogs

The decibel concept is extended to allow the ratio of any two like quantities to be expressed in decibel units, these being termed *decilogs*. For example, two temperatures T_1 and T_2 may be expressed as

$$10\log\frac{T_1}{T_2} \text{ decilogs}$$

The multiplying factor for decilogs is always 10. Decilogs may also be referred to a unit value. In practice the name decilog is seldom

used and common practice is to use decibels referred to the selected unit. For example, a temperature of 290 K (degrees kelvin) would be given as

$$10\log\frac{290\text{ K}}{1\text{ K}} \cong 24.62\text{ dBK}$$

Another example which occurs widely in practice is that of frequency referred to 1 Hz. A bandwidth of 36 MHz is equivalent to

$$10\log\frac{36 \times 10^{6}\text{ Hz}}{1\text{ Hz}} = 75.56\text{ dBHz}$$

Apart from voltage and current, all decibel-like quantities are taken as 10 log (·). In this book, brackets are used to denote decibel quantities, using the basic power definition. A ratio X in decibels is thus

$$[X] \equiv 10\log X$$

The 3-bar symbol indicates an identity. The abbreviation dB is used, with letters added where convenient to signify the reference. For example, a bandwidth $b = 3$ kHz expressed in decibels is

$$[b] = 10\log\frac{3000\text{ Hz}}{1\text{ Hz}} = 34.77\text{ dBHz}$$

In many cases the reference unit is understood and is not shown explicitly. For example, the temperature of 290 K in decilogs may appear as:

$$10\log 290 \cong 24.62\text{ dBK}$$

A quantity which occurs frequently in link-budget calculations is Boltzmann's constant, which has the dimensions of J/K:

$$k = 1.38 \times 10^{-23}\text{ J/K}$$

Expressed in decibels relative to 1 J/K, this is

$$10\log k = -228.6\text{ dB}$$

Strictly, the dB units should be written as dBJ/K for decibels relative to 1 J/K. This is awkward, and it is simply shown as −228.6 dB.

If a voltage ratio V_r is to be expressed in decibels using the bracket notation, it would be written as $2[V_r]$, since $20\log V_r = 2 \times 10\log V_r = 2[V_r]$. A similar argument applies to current ratios. Thus, it must be kept in mind that the brackets are identified always with 10 log (·).

The more commonly used decibel-type abbreviations are summarized below:

dBW: decibels relative to 1 W

dBm: decibels relative to 1 mW

dBV: decibels relative to 1 V

*dB*μV: decibels relative to 1 μV

dBK: decibels relative to 1 K

dBHz: decibels relative to 1 Hz

dBb/s or *dBbps*: decibels relative to 1 bit/s

The advantage of decibel units is that they can be added directly, even though different reference units may be used. For example, if a power of 34 dBW is transmitted through a circuit which has a loss of 20 dB, the received power would be

$$[P_R] = 34 - 20 = 14 \text{ dBW}$$

In some instances, different types of ratios occur, an example being the *G/T* ratio of a receiving system described in detail in Sec. 12.6. *G* is the antenna power gain, and *T* is the system noise temperature. Now *G* by itself is dimensionless, so taking the log of this ratio requires that the unit reciprocal temperature K^{-1} be used. Thus

$$\begin{aligned}\left[\frac{G}{T}\right] &= 10\log\left(\frac{G/T}{1 \text{ K}^{-1}}\right) \\ &= 10\log G - 10\log\left(\frac{T}{1 \text{ K}}\right) \\ &= [G] - [T] \text{ dBK}^{-1}\end{aligned}$$

The units are often abbreviated to dB/K, but this must not be interpreted as decibels per kelvin. In practice [*G/T*] is often a known parameter of the system, and its value will be stated in dBK^{-1}.

The Neper

The *neper* is a logarithmic unit based on natural, or naperian, logarithms.

Originally it was defined in relation to currents on a transmission line which decayed according to an exponential law. The current magnitude may decay from a value I_1 to I_2 over some distance *x* (for example a length of transmission line) so that the ratio is

$$\frac{I_1}{I_2} = e^{-\alpha x}$$

where α is the attenuation constant. The attenuation in nepers is defined as

$$N = -\ln\frac{I_1}{I_2}$$
$$= \alpha x$$

The same ratio expressed as an attenuation in decibels would be

$$D = -20\log\frac{I_1}{I_2}$$

Since the same ratio is involved in both definitions, it follows that

$$\frac{I_1}{I_2} = e^{-N} = 10^{-D/20}$$

Thus,

$$D = 20N\log_{10}e$$
$$= 8.686N$$

It follows that for $N = 1$, $D = 8.686$, that is, one neper is equivalent to 8.686 dB.

Index

ABOUT THE AUTHOR

Dennis Roddy, Professor Emeritus of Electrical Engineering at Lakehead University in Thunder Bay, Ontario, Canada, has more than 40 years of experience in both industrial and technical education. He is also the author of *Radio and Line Transmission,* vols. 1 and 2; *Introduction to Microelectronics;* and *Microwave Technology;* and the coauthor (with John Coolen) of *Electronic Communications and Electronics*.